

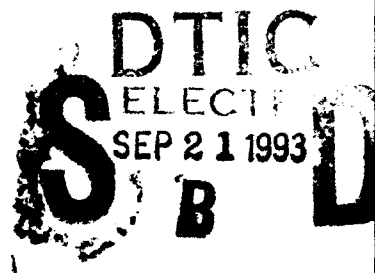
AL/HR-TP-1993-0028

AD-A269 735



**BUILDING A JOINT-SERVICE CLASSIFICATION
RESEARCH ROADMAP: CRITERION-RELATED ISSUES**

Deirdre J. Knapp
John P. Campbell



Human Resources Research Organization (HumRRO)
66 Canal Center Plaza, Suite 400
Alexandria, VA 22314

**HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL RESEARCH DIVISION
7909 Lindbergh Drive
Brooks Air Force Base, TX 78235-5352**

July 1993

Final Technical Paper for Period January 1992 - April 1993

Approved for public release; distribution is unlimited.

93-21871



93 3 20 06 5

**AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS**

ARMSTRONG

LABORATORY

NOTICES

This technical report is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this paper does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

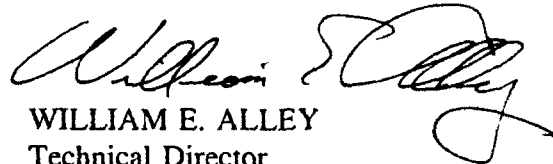
When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.


This report has been reviewed and is approved for publication.



MALCOLM JAMES REE
Contract Monitor



WILLIAM E. ALLEY
Technical Director
Manpower & Personnel Research Div


Acting
ROGER W. ALFORD, Lt Colonel, USAF
Chief, Manpower & Personnel Research Div

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
<small>Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.</small>				
1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE July 1993	3. REPORT TYPE AND DATES COVERED Final January 1992 - April 1993		
4. TITLE AND SUBTITLE Building a Joint-Service Classification Research Roadmap: Criterion-Related Issues		5. FUNDING NUMBERS C - F33615-91-C-0015 PE - 62205F PR - 7719 TA - 25 WU - 01		
6. AUTHOR(S) Deirdre J. Knapp John P. Campbell				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization (HumRRO) 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314		8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) Armstrong Laboratory (AFMC) Human Resources Directorate Manpower and Personnel Research Division 7909 Lindbergh Drive Brooks Air Force Base, TX 78235-5352		10. SPONSORING / MONITORING AGENCY REPORT NUMBER AL/HR-TP-1993-0028		
11. SUPPLEMENTARY NOTES Armstrong Laboratory Technical Monitor: Malcolm James Ree, (210) 536-3942				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.		12b. DISTRIBUTION CODE		
13. ABSTRACT (Maximum 200 words) The Air Force Armstrong Laboratory, the Army Research Institute, the Navy Personnel Research and Development Center, and the Center for Naval Analyses are committed to enhancing the overall efficiency of the Services' selection and classification research agenda. This means reducing redundancy of research efforts across Services and improving inter-Service research planning, while ensuring that each Service's priority needs are served. The Roadmap project is composed of six tasks. The first task identified classification research objectives. Tasks 2 through 5 consist of reviews of specific predictor, job analytic, criterion, and methodological needs of each of the methods and issues as they relate to the selection and classification research objectives outlined in Task 1 of the Roadmap project (Russell, Knapp, & Campbell, 1992).				
14. SUBJECT TERMS Criterion measures Criterion related Roadmap		Validity		15. NUMBER OF PAGES 102
				16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UL	

**BUILDING A JOINT-SERVICE CLASSIFICATION RESEARCH
ROADMAP: CRITERION-RELATED ISSUES**

TABLE OF CONTENTS

	<u>Page</u>
CHAPTER I: INTRODUCTION	1
Scope of Report	2
Overview of Report	3
CHAPTER II: THE JOINT-SERVICE JOB PERFORMANCE MEASUREMENT PROJECT	5
The Air Force JPM Project	6
The Army JPM Project	8
The Navy JPM Project	12
The Marine Corps JPM Project	14
Summary	16
CHAPTER III: A GENERAL JOB PERFORMANCE MODEL	18
Performance Defined	18
Determinants of Performance	18
Latent Structure of Performance	19
Analysis of the JPM Project's Coverage of Performance	21
CHAPTER IV: MEASUREMENT METHODS AND EVALUATION FACTORS	24
Criterion Measurement Methods	24
Characteristics of High Quality Criteria	25
Format for Criterion Measure Review	28
CHAPTER V: PERFORMANCE TESTS	30
Work Samples	30
Content Sampling	30
Construction	35
Scorer Qualifications and Training	39
Test Administration	40
Scoring	40
Evaluation	41
Simulations	44
Computer/Visual/Audio Aids	45
Assessment Center Exercises	49

TABLE OF CONTENTS (Continued)

	<u>Page</u>
CHAPTER VI: VERBAL TESTS	52
Types of Verbal Tests	52
Structured Response Format	52
Unstructured Response Format	54
Evaluation of Verbal Tests	55
Summary	57
CHAPTER VII: PERFORMANCE RATINGS	58
Summary of Ratings Collected in JPM Project	58
Rating Sources	59
Rating Scale Content	60
Evaluation of Performance Ratings	62
Further Analysis of JPM Data	65
Summary	66
CHAPTER VIII: ARCHIVAL RECORDS	67
Supervisor Ratings	67
Promotion Rate	68
Training Grades	68
Personnel File Records	68
Production Indices	69
Attrition/Turnover	70
Summary	71
CHAPTER IX: DISCUSSION AND CONCLUSIONS	72
Choosing Among Alternative Criterion Measures: Conceptual Issues	72
Research vs. Appraisal	72
The General Research Objective	72
The Specific Goals of Selection and Classification	75
Choosing Among Alternative Criterion Measures: Practical Issues	76
The Issue of Measurement Bias	77
Individual Contributions to Team Performance	78
Utility of JPM Instruments and Data	78
Revisiting the Roadmap Objectives	80
REFERENCES	81

TABLE OF CONTENTS (Continued)

Page

LIST OF TABLES AND FIGURES

Tables

Table 1 -	Uncorrected Correlations Between AFQT and Air Force Criterion Factors	7
Table 2 -	Air Force Correlations Between Performance Tests and AFQT ...	7
Table 3 -	Project A/Career Force Military Occupational Specialties (MOS) ..	8
Table 4 -	Project A/Career Force Sample Sizes	11
Table 5 -	Mean Incremental Validity for the Composite Scores Within Each Predictor Domain	13
Table 6 -	Correlations Between Predictor Constructs and Core Technical Proficiency	13
Table 7 -	Navy JPM Project Ratings	14

Figures

Figure 1	Project A/Career Force Research Flow and Samples	9
Figure 2	Summary of Data Collected in the Joint-Service JPM Project	17
Figure 3	Determinants of Job Performance Components (PC)	19
Figure 4	Summary of Measurement Methods	29
Figure 5	Sample Air Force Hands-On Test	36
Figure 6	Army Supervisory Role-Play Scenarios	50
Figure 7	Army Situational Judgment Test Sample Item	53
Figure 8	Performance Ratings Collected in Joint-Service JPM Project	59
Figure 9	Force-Wide Rating Dimensions Used in JPM Project	63

DTIC TAB

Accession For	
NTIS GRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

PREFACE

This effort was conducted as part of work unit 77192403 to investigate issues in the selection and classification of Air Force Military personnel. The authors thank Malcolm Ree who monitored the contract for support and guidance.

It is hoped that this effort will serve as a guide to future research and development to bring about improvements in military manpower systems, especially in light of decreasing resources.

BUILDING A JOINT-SERVICE CLASSIFICATION RESEARCH ROADMAP: CRITERION-RELATED ISSUES

I INTRODUCTION

The Air Force Armstrong Laboratory, the Army Research Institute for the Behavioral and Social Sciences, the Navy Personnel Research and Development Center, and the Center for Naval Analyses are committed to enhancing the overall efficiency of the Services' selection and classification research agenda. This means reducing redundancy of research efforts across Services and improving inter-Service research planning, while ensuring that each Service's priority needs are served. With these goals in mind, the Armstrong Laboratory and the Army Research Institute have contracted with the Human Resources Research Organization (HumRRO) to develop a Joint-Service selection and classification research Roadmap.

The Roadmap project has six tasks. The first task, Identify Classification Research Objectives, involved interviews with selection and classification experts and decision-makers from each Service to determine research objectives. Tasks 2 through 5 consist of reviews of specific predictor, job analytic, criterion, and methodological needs of each of the Services. The final task, Prepare a Research Roadmap, will integrate findings of Tasks 1 through 5 into a master research plan.

This report documents the findings of Task 4. The goal of Task 4 is to review and discuss criterion measurement methods and issues as they relate to the selection and classification research objectives outlined in Task 1 of the Roadmap project (Russell, Knapp, & J. P. Campbell, 1992). The objectives identified in Task 1 which most directly express the Services' interest in criterion measurement issues are the following:

- Investigate criterion issues (e.g., How does the type of criterion used in validation affect estimates of classification efficiency and, ultimately, classification decisions? What is the appropriate criterion?).
- Investigate alternative selection and classification criterion measures in terms of their relative construct validity and susceptibility to subgroup bias.

Other selection and classification objectives identified in Task 1 which are related, albeit more tangentially, to criterion measurement issues are as follows:

- Investigate ways to maximize the influence of predicted performance in the assignment system (e.g., improve composite standard setting procedures; incorporate predicted performance into assignment algorithm).
- Improve classification efficiency by improving strategies to generalize classification research findings across jobs and military populations.

- Develop and evaluate alternative strategies and models for estimating the cost-effectiveness of an alternative classification system in terms of reduced training costs, reduced attrition, dollars, etc.

We will not attempt to satisfy these selection and classification objectives in this report. Rather, we will review criterion measurement research and related issues, then close the report with a revised set of objectives which clarify and broaden the objectives listed above.

Scope of Report

As we discuss further in Chapter II, the research literature on criterion measurement is smaller than one would expect given its central importance in the field of industrial and organizational psychology. Even so, there is much more research than we will be able to fully address. Instead, the relative emphasis we have chosen to give the discussion of various criterion measurement issues and measurement methods reflects our judgment of their potential usefulness for advancing military selection and classification.

An organization's selection and classification system may be designed to serve any number of goals. The most common purpose of such a system is to maximize job performance levels by identifying those applicants who are likely to be more or less able to perform the job. Additional or alternate goals may be to predict how well individuals will perform in job training, how satisfied they will be with their jobs, or how long they can be expected to stay with the organization (i.e., turnover). Historically, the military's central classification goal has been the prediction of performance in advanced technical training. This approach reflects the theory that learning is a prerequisite for success on the job, a theory that has been substantially supported (e.g., Hunter, 1986). Over the last decade, however, the overriding goal has become the prediction of on-the-job performance, itself. This priority is reflected in the content of this report which emphasizes the conceptualization and measurement of job performance. We discuss training performance as a potential surrogate measure of job performance, and we do not address job satisfaction at all. This is not to say that these are not legitimate organizational concerns. However, we agree with the philosophy that the prediction of on-the-job performance should be the paramount goal, particularly when evaluating the fairness and legality of a selection and classification system.

Because of the great cost of attrition to the military, and the fact that it was raised as an issue in the research objectives identified in Task 1, we discuss turnover briefly in this report as well. We will also examine other administrative indices of performance (e.g., commander's ratings, awards received) because such indices may provide valid measures of some aspects of performance which will be available for all or most personnel at minimal cost. This is especially important because it is unlikely that for-research-only measures could routinely be collected on large numbers of job incumbents for most jobs in the military. In fact, we will assume that this is a requirement that will have to be met indirectly with the assistance of technical approaches such as validity generalization and synthetic validity.

Most of the content of this report is generalizable to all types of jobs in the work force (e.g., civilian and military, enlisted and officer). To the extent that we discuss specific measurement needs in the military, however, we will confine the discussion to enlisted positions. Furthermore, a significant portion of this report's content is pertinent for performance measurement needs in general, regardless of whether they are intended to serve research or operational goals. To the extent that the requirements for research and operational needs diverge, however, our discussion will address the requirements for research measures.

Finally, the reader will find that the vast majority of research presented in this report was conducted by the military. This can be largely explained by the fact that this is what is available. The Services have been, and continue to be, at the forefront of criterion measurement research. In contrast, there appears to be a relative sparsity of such research in the civilian sector. It may well be that civilian organizations are less willing to devote resources to this area of research. It is probably also true, however, that they are less likely than military organizations to publish their findings. That is, much of their research may simply not be generally available. In part, this may be explained by the fact that it is common practice in the civilian sector to rely on content validity-based evidence to support the use of selection and classification, promotion, certification, and licensing procedures. Thus, there is a great deal of effort to develop quality tests, but because the tests are used for decision-making purposes (e.g., selection, promotion, certification), users are probably reluctant to publish details about them. Also, the operational nature of most of this work is such that it is not publishable in most of our research journals. That is, there are no evaluations, experimental controls, or other innovations which justify research publication. Educational and Psychological Measurement is a notable exception to this phenomenon in that it publishes "simple" validity studies. Few of these studies, however, are concerned with the criterion-related validation of personnel selection or classification measures.

A second, and equally important, reason for our focus on military research is that we believe that a concise overview and discussion of the military's recent efforts in the area of performance measurement is necessary for achieving the intended goals of this research Roadmap effort. Wigdor and Green (1991) provide a review and summary of the Service's recent efforts on performance measurement. Our review complements their work by adding more up-to-date information and providing greater detail regarding specific criterion measures.

Overview of Report

The next three chapters will be used to provide the reader with multiple, though related, contexts for the consideration of criterion measurement issues in the military. Over the past decade, the Services have devoted considerable resources to an extensive job performance measurement project. Chapter II provides an overview of each of the Service's contributions to this research. Chapter III then describes a conceptual model of performance, and Chapter IV presents a taxonomy of measurement methods and a review of factors that can be used to evaluate criterion measures. Taken together, these three chapters will provide the foundation for examining individual measurement

methods and related issues, which is done in Chapters V through VIII, and for the final review and commentary which comprises the final chapter of the report.

II. THE JOINT-SERVICE JOB PERFORMANCE MEASUREMENT PROJECT

In spite of the eloquent pleas of some very well-known, applied psychologists (Dunnette, 1963; Jenkins, 1946; Wallace, 1965), the criterion has received the least attention of any of the parameters in various models of personnel selection and classification. Decades of research have been devoted to the nature of abilities, the structure of personality, and the assessment of interests (i.e., the predictors). Similarly, examinations of the statistical and psychometric properties of the joint predictor and criterion distribution have filled the research literature. By comparison, the literature pertaining to the structure and content of performance is relatively barren. Given the crucial role of performance in virtually all research on personnel decisions, this is indeed a strange situation.

The Services took a large step forward in addressing the sparsity of research focusing on the "criterion problem" when they embarked upon the Joint-Service Job Performance Measurement/Enlistment Standards (JPM) project (D. A. Harris, 1987). The JPM project was initiated as a result of a 1980 Congressional mandate which directed the Services to demonstrate empirically that performance on the Armed Services Vocational Aptitude Battery (ASVAB) is predictive of performance on the job. Previously, evidence of ASVAB's predictive validity had been based on training performance criterion measures. As part of the JPM project initiative, each of the Services embarked on individual programs of performance measurement research which were coordinated through the Joint-Service Job Performance Measurement (JPM) working group. Independent technical oversight of the efforts was provided by the National Academy of Science (NAS) Committee on the Performance of Military Personnel (e.g., Wigdor & Green, 1986).

At the outset of the performance measurement project, the JPM working group identified hands-on work sample tests as the benchmark job performance measurement method against which less costly "surrogate" measurement methods, such as performance ratings or written job knowledge tests would be compared. Furthermore, each Service was tasked to measure job performance for a sample of jobs using hands-on tests as well as a least one surrogate measure. Surrogate measurement methods were assigned to each of the Services in an effort to reduce redundancy of research resources. The surrogate assignments were as follows: performance-based interview testing for the Air Force, written job knowledge testing for the Army, simulator testing for the Navy, and operational indices (e.g., training school scores) for the Marine Corps.

The following section provides a summary of each of the Services' JPM research programs, and their respective approaches to job performance measurement. Even though the JPM working group adopted a common, coordinated strategy, the various approaches to performance measurement implemented by each of the Services illustrate their individual views of what constitutes "job performance." Information provided includes the types of jobs studied, the job analysis procedures used to support criterion development, the types of criterion measures used, the basic validation research design, and a brief summary of selected validation results. A more thorough discussion of the

critterion measures developed by each of the Services in this research program is provided in subsequent chapters.

The Air Force JPM Project

In the Air Force JPM project, job performance measures were developed for eight Air Force Specialties (AFS). Two jobs were selected from each of the four job groupings used for classification in the Air Force (mechanical, administrative, general, and electronic). The sampled jobs were:

- Jet Engine Mechanic (AFS 426X2)
- Information Systems Radio Operator (AFS 492X1)
- Air Traffic Control Operator (AFS 272X0)
- Avionic Communications Specialist (AFS 328X0)
- Aerospace Ground Equipment Specialist (AFS 423X5)
- Personnel Specialist (AFS 732X0)
- Aircrew Life Support Specialist (AFS 122X0)
- Precision Measuring Equipment Specialist (AFS 324X0)

The Air Force developed work samples, interview tests, and behaviorally-based rating instruments for each of the eight jobs (Hedge & Teachout, 1986). For the second subset of four jobs listed above, written job knowledge tests were also developed. Finally, archival training school scores were extracted from existing Air Force records.

With the exception of one AFS, data were collected at 13 Air Force bases within the continental United States. Data for the Information Systems Radio Operators were collected world-wide. Examinees, each of whom had 13-48 months time-in-service, were selected randomly at each of the bases where testing took place. Approximately six to ten hours, depending upon AFS, were required to administer the full set of criterion measures. Instruments were developed and data were collected over a five-year period, with data for the first AFS (Jet Engine Mechanics) being collected early in 1985.

AFS-specific validation results have been published for only the first four AFS that were studied. Exploratory factor analysis of these four AFS yielded five performance factors (Office of the Assistant Secretary of Defense (Force Management and Personnel) [OASD (FM&P)], 1989):

- Technical proficiency
- Interpersonal proficiency
- Supervisor ratings
- Self ratings
- Peer ratings

Validity analyses suggest that the Armed Forces Qualification Test (AFQT) composite of the ASVAB is quite variable across AFS in its ability to predict performance on these factors (see Table 1), although a meta-analytic cumulation of these results would likely reduce the magnitude of this apparent variability. AFQT was moderately to highly

predictive of all five factors for one AFS and weakly predictive of only one or two factors for two other AFS. Technical proficiency was the only factor which consistently showed a relationship to AFQT scores.

Table 1				
Uncorrected Correlations Between AFQT and Air Force Criterion Factors				
	Jet Engine Mechanic (n=255)	Info. Sys. Operator (n=156)	Air Traffic Control (n=172)	Avionic Comm. (n=98)
Technical Proficiency	.18	.37	.11	.33
Interpersonal Proficiency	.08	.19	-.04	.14
Supervisor Ratings	.04	.28	.04	.18
Self Ratings	.04	.16	-.09	-.02
Peer Ratings	.11	.36	.04	.14

Note. Adapted from Joint-Service efforts to link enlistment standards to job performance. Recruit quality and military readiness by the Office of the Assistant Secretary of Defense (Force Management and Personnel), 1989.

With regard to analyses which focus on hands-on and interview criteria, hands-on tests correlated .23 (uncorrected) and interview tests correlated .21 with AFQT across all eight AFS (Hedge & Teachout, 1992). Corresponding results, by AFS, are noted in Table 2. Again, there is considerable variability in results across AFS. Comparison of results obtained using hands-on and interview scores are discussed further in Chapter V.

Table 2		
Air Force Correlations Between Performance Tests and AFQT		
	Hands-on	Interview
Jet Engine Mechanic (n=255)	.07	.19
Information Systems Radio Operator (n=156)	.32	.36
Air Traffic Control Operator (n=172)	.10	.16
Avionic Communications Specialist (n=98)	.28	.33
Aerospace Ground Equipment Specialist (n=264)	.18	.08
Personnel Specialist (n=218)	.27	.24
Aircrew Life Support Specialist (n=216)	.11	.08
Precision Measuring Equipment Specialist (n=138)	.28	.23

Note. Adapted from Interview testing as a work sample measure of job proficiency by J. W. Hedge, M. S. Teachout, and F. J. Laue, 1990, AFHRL-TP-90-61, Brooks AFB, TX: Air Forces Human Resources Laboratory.

The Army JPM Project

Two related projects comprise the foundation of the Army's JPM research program: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel (known as Project A) and Building and Maintaining the Career Force (known simply as Career Force). The Project A/Career Force research program was designed to develop a valid and comprehensive set of performance criterion measures as well as to develop an array of experimental selection and classification instruments to validate in conjunction with validation of the ASVAB.

The project included nine "Batch A" MOS for which a full array of criterion measures was developed and ten "Batch Z" MOS for which a subset of criterion measures was developed (J. P. Campbell & Zook, 1990b). The Batch A and Batch Z MOS are listed in Table 3. Figure 1 graphically depicts the Project A/Career Force research plan and provides a nomenclature for referring to the various data collection samples. The first major data collection took place in 1985, and was a concurrent validation study with a sample size of approximately 9,500. The predictor and criterion measures developed for the concurrent validation were refined and used in a longitudinal study which began in 1986. The longitudinal validation included a series of data collections to gather predictor (1986-1987), training (1987), first tour performance (1988), and second tour (1991-1992) performance data. A major try-out of the second tour measures occurred in conjunction with the first tour longitudinal data collection.

Table 3			
Project A/Career Force Military Occupational Specialties (MOS)			
Batch A MOS		Batch Z MOS	
11B	Infantryman	12B	Combat Engineer
13B	Cannon Crewman	16S	MANPADS Crewman
19E	M1 Armor Crewman	27E	TOW/Dragon Repairer
19K	M1A1 Armor Crewman	29E	Electronics Repairer
31C	Single Channel Radio Operator	51B	Carpentry/Masonry Specialist
63B	Light Wheeled Vehicle Mechanic	54B	Chemical Operations Specialist
71L	Administrative Specialist	55B	Ammunition Specialist
88M	Motor Transport Operator	67N	Utility Helicopter Repairer
91A	Medical Specialist	76W	Petroleum Supply Specialist
95B	Military Police	76Y	Unit Supply Specialist
		94B	Food Service Specialist
		96B	Intelligence Analyst

Note. 19E incumbents generally transitioned to 19K as M1A1 tanks were fielded during the course of the research; 76W was not included in the longitudinal validation; 29E and 96B were not included in the concurrent validation; Adapted from Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel edited by J. P. Campbell and L. M. Zook, 1990, ARI-TR-952, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences.

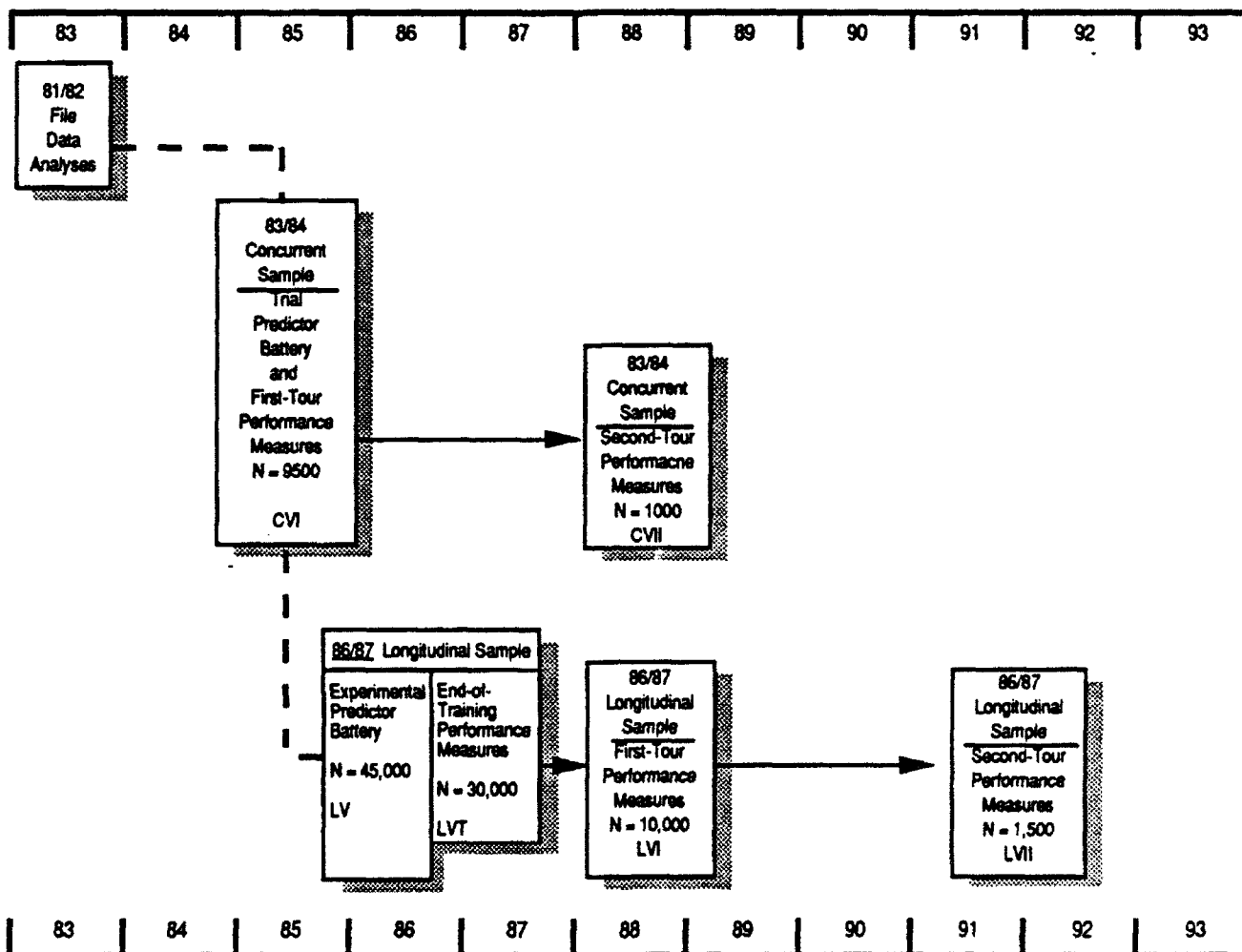


Figure 1. Project A/Career Force Research flow and samples.

Note. Adapted from Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A edited by J.P. Campbell and L.M. Zook, 1990, ARI-RR-1597, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences.

Data from the concurrent validation have been used in all of the joint-Service JPM analyses. In this data collection, soldiers were tested at 13 Army posts throughout the continental United States and at numerous locations throughout Germany. For smaller density MOS, all soldiers at a given location who were available for testing were tested. For larger density MOS, a random sampling plan was followed as much as possible given the constraints of each installation.. All soldiers selected for testing entered the service between 1 July 1983 and 30 July 1984 (i.e., they had approximately 18-24 months time-in-service).

For the longitudinal validation, experimental predictors were administered to new recruits who entered one of the sampled MOS during 1986-1987. No sampling plan was necessary because every effort was made to test all new recruits who in-processed into one of the sampled MOS during the time period that predictors were being administered. Figure 1 provides approximate usable sample sizes for each of the major data collections. MOS-specific sample sizes are listed in Table 4. As with the concurrent validation, data were collected world-wide (including the Republic of South Korea for the second tour sample).

The approach to performance measurement used in Project A/Career Force was based on the assumption that job performance is multidimensional and has a latent structure which can be identified empirically. Multiple measurement methods were used to capture the job requirements identified through job analysis. One day of criterion assessment was developed for each MOS. The measurement methods used can be summarized as follows:

- Hands-on tests
- Supervisory role-play exercises (Second tour only)
- Written job knowledge tests
- Written training knowledge tests
- Written supervisory situational judgment test (Second tour only)
- Administrative criteria (e.g., separation status, promotion rate)
- Behavioral ratings

An iterative process involving content analysis, principal components analysis, and, finally, confirmatory factor analysis using data from the first tour concurrent (J. P. Campbell, McHenry, & Wise, 1990) and longitudinal validation (Childs, Oppler, & Peterson, 1992) samples yielded the most support for a model of performance that consisted of five substantive factors and two methods factors:

- (1) General soldiering proficiency
 - (2) Core technical proficiency
 - (3) Effort and leadership
 - (4) Personal discipline
 - (5) Physical fitness and military bearing
-
- (1) Ratings method
 - (2) Written method

Table 4						
Project A/Career Force Sample Sizes						
	Concurrent Validation		Longitudinal Validation			
	CV	CVII	Predictor	Training	First Tour	Second Tour
Unknown			902			
11B	702	127	14193	8117	909	346
13B	667	162	5087	4712	916	179
19E	503	33	583	442	249	
19K		10	1849	1606	824	168
31C	366	103	1072	667	529	71
63B	637	116	2241	1215	752	194
71L	686	112	2140	1414	678	155
88M	514	144	1593	1354	682	89
91A	501	105	4219	3218	824	221
95B	692	141	4206	3639	452	168
12B	704		2118	1872	841	
16S	470		800	585	472	
27E	147		139	92	90	
29E			257	138	112	
51B	108		455	353	213	
54B	434		967	616	499	
55B	291		482	389	279	
67N	276		334	233	197	
76W	490					
76Y	630		2756	1651	788	
94B	612		3522	1806	832	
96B			320	196	128	
Total	9430	1053	50235	34315	11266	1591

Note. Second tour longitudinal validation sample sizes are preliminary counts; First tour sample includes some soldiers who were not tested on the experimental predictors; Information from Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A edited by J. P. Campbell and L. M. Zook, 1990, ARI-RR-1597, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences and D. Steele, personal communication, 21 August 1992.

A comparable solution has been identified for second tour performance. The best fitting model was used to assign individual measures to the performance factors and scores for each dimension were obtained by a simple sum of standard scores. These performance factor scores have been used as the principal criteria in the project's validation analyses. For the sake of completeness, analyses were also run using a number of individual criterion measures as well.

The J. P. Campbell et al. (1990b) report illustrates what can be gained from criterion analyses that use structural equation modeling procedures and confirmatory techniques. For the Project A data it was possible to control for the obfuscating effects of method variance by postulating two orthogonal methods factors and then making substantive hypotheses about the covariance structure of the residuals. Evidence for the real world meaning of the substantive factors, even when method variance is not partialled from the simple sum factor score, is found in their differential correlations with other variables. An exploratory factor analysis would not have modeled performance in the same way and would have missed some very meaningful features of the latent structure. Too often, using exploratory factor analysis, methods define factors, or the general factor is taken to be the only meaningful one. This point is also illustrated by the reanalysis of some of the Air Force JPM data by Vance et al. (1988) using confirmatory techniques. By postulating uncorrelated methods factors corresponding to type of rater, it was possible to test for substantive performance factors.

Table 5 summarizes the validity findings based on the concurrent validation data (McHenry, Hough, Toquam, Hanson, & Ashworth, 1990). Results indicate that the first two "can-do" performance factors are predicted well with general cognitive ability (i.e., ASVAB scores), and other experimental predictor constructs add little incremental validity. The latter three "will-do" factors, however, are less well predicted by general cognitive ability but the temperament/personality measure provides considerable incremental validity over ASVAB used alone. Similar results have been found using the first tour longitudinal validation data (Oppler & Peterson, 1992). To provide an example of MOS differences in observed validity estimates, Table 6 shows estimates for the prediction of core technical proficiency using four different ASVAB factor scores. The Cannon Crewman MOS (13B) stands out as the job least well predicted by the ASVAB.

The Navy JPM Project

Four Navy ratings (i.e., jobs), each of which uses a different ASVAB composite for classification, comprised the major focus of the Navy's JPM research program: Machinist's Mate, Radioman, Electronics Technician, and Electrician's mate. Limited data collections were conducted for Gas Turbine Mechanics and Fire Control Technicians (Van Hemel, Alley, Baker, & Swirski, 1990). In addition, performance measures developed by the Air Force for J-79 Jet Engine Mechanics (AFS 426X2) were adapted for use with Navy/Marine Corps Aviation Machinist's Mates in an effort to demonstrate the utility of cross-Service transfer-of-technology possibilities (Baker & Blackhurst, 1986).

Table 5						
Mean Incremental Validity ^{a, b} for the Composite Scores Within Each Predictor Domain						
		Predictor Domain				
Job Performance Factor	General Cognitive Ability (K=4) ^c	General Cognitive Ability Plus Spatial Ability (K=5)	General Cognitive Ability Plus Perceptual- Psychomotor Ability (K=10)	General Cognitive Ability Plus Temperament/ Personality (K=8)	General Cognitive Ability Plus Vocational Interest (K=10)	General Cognitive Ability Plus Job Reward Preference (K=7)
Core Technical Proficiency	.63	.65	.64	.63	.64	.63
General Soldering	.65	.68	.67	.66	.66	.66
Effort and Leadership	.31	.32	.32	.42	.35	.33
Personal Discipline	.16	.17	.17	.35	.19	.19
Physical Fitness and Military Bearing	.20	.22	.22	.41	.24	.22

^aValidity coefficients were corrected for range restriction and adjusted for shrinkage.

^bIncremental validity refers to the increase in R afforded by the new predictors above and beyond the R for the Army's current predictor battery, the ASVAB.

^cK is the number of predictor scores.

Table 6									
Correlations Between Predictor Constructs and Core Technical Proficiency ^a									
ASVAB Factors	11B	13B	19E	31C	63B	88M	71L	91A	95B
Technical	.60	.36	.56	.59	.69	.55	.37	.61	.51
Verbal	.63	.33	.49	.67	.50	.44	.56	.71	.59
Quantitative	.60	.32	.49	.67	.45	.46	.63	.64	.59
Speed	.48	.25	.28	.57	.29	.27	.52	.56	.47

^aCorrected for range restriction.

Note. Tables 5 and 6 from Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year edited by J. P. Campbell, 1988, ARI-RR-792, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences.

For each rating in the primary JPM sample, the Navy developed hands-on tests as well as one or more surrogate performance measures. The surrogate measures were simulations (using computer-aided, video-based, and/or paper-and-pencil technologies) and, for two ratings, behaviorally-based performance ratings. Table 7 shows the types of measures developed for each rating and their respective sample sizes.

Table 7		
Navy JPM Project Ratings		
Machinist's Mate	n=184	Hands-on, Written, Ratings
Radioman	n=260	Hands-on, Written, Ratings
Electronics Technician	n=136	Hands-on, Interactive video, TRT
Electrician's Mate	n= 80	Hands-on, Interactive video

Note. From G. Laabs, personal communication, 22 September 1992; TRT refers to a video-based Tailored Response Test.

Although Radioman testing was restricted to three test sites, JPM testing for the remaining ratings took place in as many as 27 or more locations (primarily on ship) (G. Laabs, personal communication, 22 September 1992). Testing required approximately 10 hours, with hands-on testing taking roughly four to six hours of that time. The remaining time was for administration of surrogate measures (e.g., written tests, computerized tests), rating scales, and experimental ability-based predictors developed by the Navy (Wolfe, Alderton, Cory, & Larson, 1987).

Few validation results using the Navy JPM data have been reported. Kroecker and Bearden (1987), however, presented validation results for the Machinist's Mate sample. Tasks were grouped into three functional categories for purposes of analysis. Validity estimates for ASVAB subtests predicting performance in these three categories ranged from .025 to .425 for the hands-on tests, -.025 to .375 for the written job knowledge tests, -.125 to .275 for self ratings, -.125 to .325 for peer ratings, and -.075 to .325 for supervisor ratings were no statistical corrections.

The Marine Corps JPM Project

In the early 1980's, the Marine Corps conducted a preliminary criterion measurement study using three MOS (Maier & Hiatt, 1985). Following that study, the Marine Corps constructed a full-scale JPM research plan which called for the inclusion of at least two MOS from each of the four ASVAB composite areas used for classification (i.e., General Technical (GT), Mechanical Maintenance (MM), Electronics (EL), and Clerical (CL)). Budget cuts led to the premature end of the research effort in 1992, however, before research was conducted on MOS in the electronics and clerical occupations. The Marine Corps JPM effort, then, included five MOS from the Infantry

Occupational Field (GT) and two sets of MOS from the Mechanical Maintenance Occupational Field (MM):

- Infantry Rifleman (n=993)
- Machinegunner (n=300)
- Mortarman (n=277)
- Assaultman (n=310)
- Careerist Infantry Unit Leader (n=387)

- CH-46, CH-53A/D, U/AH-1, and CH-53E Helicopter Mechanics (n=559)
- Automotive Mechanic (n=891)

As with the Army, the Marine Corps was interested in evaluating the ability to predict performance beyond the first term. For this reason, they included the infantry unit leader MOS in the research. The sample sizes provided above are approximate, having been drawn from several sources providing validity results (e.g., OASD, 1989).

Hands-on performance tests, written job knowledge tests, and supervisor rating scales were developed for each job in the Marine Corps sample. In addition, training school grades and supervisor ratings of performance were retrieved from Marine Corps records for inclusion in the research (Carey, 1990; Crafts et al., 1991).

JPM testing took place at two test sites. These sites were selected because the preponderance of Marines in each of the tested MOS were stationed there. Marines in those units available for testing were selected for participation in the JPM project using a stratified random sampling strategy (Mayberry, 1988). The stratification variables were pay grade, educational level, and time-in-service.

Testing occurred over two days. One day was devoted to hands-on testing and the other day was devoted to the administration of written knowledge tests and experimental predictor tests from the Joint-Service Enhanced Computer Assisted Test (ECAT) battery. Infantry MOS testing took place in 1987 and Mechanical Maintenance MOS testing took place in 1990.

Validation analyses using the hands-on test scores as the criterion yielded reasonably high predictive validity estimates. Validity estimates corrected for range restriction ranged from .48 (Mortarman) to .68 (Machinegunner) for the enlistment GT composite in the Infantry MOS (N.B. Carey, personal communication, 23 April 1992). Corresponding validity estimates using the MM composite on the Mechanical Maintenance MOS ranged from .35 (CH-53E(B) Helicopter Mechanic) to .70 (CH-46 Helicopter Mechanic). The .35 estimate is a bit misleading in that the remaining estimates were all .68 or higher.

One approach used by the Marine Corps to evaluate the usefulness of the various surrogate criterion measures was to calculate validity estimates using them as well. Using the Infantry MOS in combination, the following results using the GT composite were reported by Carey (1990):

Hands-on	.69
Core job knowledge	.77
Training GPA	.52
Video firing	.46
Proficiency ratings	.31
Conduct ratings	.22
Supervisor ratings	.19

Note that these validity estimates have been corrected for range restriction in the predictor but not for criterion unreliability. Thus, some of the variation in these estimates is probably attributable to differences in criterion reliability.

Summary

The Service's have produced a massive amount of criterion measurement information over the past decade. The JPM data sets are large, both in terms of sample sizes and types of variables, and cover a wide variety of jobs. Figure 2 attempts to summarize some of this information in an abbreviated fashion. In addition to the measures and data generated, the lessons to be learned from the development, administration, and analysis of these measures are significant.

It is difficult to directly compare the validation results of the different Services because of differences in analytical approaches and incompleteness of reported results. The most critical question, whether or not ASVAB is predictive of job performance as measured by hands-on tests, however, appears to have been answered in the affirmative by all of the Services (OASD, 1991). Comparing validity levels, however, or comparing answers to other research questions (e.g., the latent structure of performance in each job) is more problematic because of the different research and analytic strategies used by each of the Services. For example, researchers conducting the DoD Linkage project have input hands-on data from each of the Services into a common analysis designed to model cost-performance trade-offs (D. A. Harris et al., 1991). A major issue they had to tackle was how to scale the data such that scores would be reasonably comparable across jobs and across Services.

Although differences in JPM measurement approaches cannot be changed because the data are already collected, further cooperative research among the Services in which comparable data analysis strategies are used is feasible. These cooperative analyses could offer much more powerful tests of various research hypotheses than that which could be achieved using the data of one Service (or one job) by itself.

	n	Hands-On	Simulation	Written	Ratings	Archival
Air Force						
Jet Engine Mechanic	255	x			x	x
Info. Systems Radio Operator	156	x			x	x
Air Traffic Control Operator	172	x			x	x
Avionic Communications Specialist	98	x			x	x
Aerospace Ground Equipment Specialist	264	x		x	x	x
Personnel Specialist	218	x		x	x	x
Aircrew Life Support Specialist	216	x		x	x	x
Precision Measuring Equipment Specialist	138	x		x	x	x
Army (Batch A, Concurrent Validation sample sizes only)						
Infantryman	702	x	x	x	x	x
Cannon Crewman	667	x		x	x	x
M1/M1A1 Armor Crewman	503	x		x	x	x
Single Channel Radio Operator	366	x		x	x	x
Light Wheeled Vehicle Mechanic	637	x		x	x	x
Administrative Specialist	686	x		x	x	x
Motor Transport Operator	514	x		x	x	x
Medical Specialist	501	x		x	x	x
Military Police	692	x	x	x	x	x
Navy						
Machinist's Mate	184	x		x	x	x
Radioman	260	x		x	x	x
Electronics Technician	136	x	x			
Electrician's Mate	80	x	x			
Gas Turbine Mechanic	88	x				
Fire Control Technician	--	x				
Marine Corps						
Infantry Rifleman	993	x	x	x	x	x
Machinegunner	300	x	x	x	x	x
Mortarman	277	x	x	x	x	x
Assaultman	310	x	x	x	x	x
Career Infantry Leader	387	x	x	x		
Helicopter Mechanics (4 jobs)	559	x		x	x	x
Automotive Mechanic	891	x		x	x	x

Figure 2. Summary of Data Collected in the Joint-Service JPM Project.

III. A GENERAL JOB PERFORMANCE MODEL

In this chapter, we describe a general model of performance which will be used as a vehicle to compare and contrast the Services' approaches to performance measurement, and which will frame subsequent discussion of individual measurement methods. This model was introduced by Campbell (1990b) and elaborated upon further by J. P. Campbell, McCloy, Oppler, and Sager (1992).

Performance Defined

The definition of performance in this model is meant to be consistent with Binning and Barrett (1989), Dunnette (1963), J. P. Campbell, Dunnette, Lawler, and Weick (1970), J. P. Campbell and R. J. Campbell (1988), and Wallace (1965). It is also consistent with the measurement models explicitly adopted by the Air Force (Kavanaugh, Borman, Hedge, & Gould, 1986) and the Army (J. P. Campbell, 1987) in their JPM research efforts, as well as the framework proposed by Murphy (1987) in JPM work conducted for the Navy.

Performance is defined as behavior or action. It is what people do. Performance includes only those behaviors or actions that are relevant to the organization's goals. Note that the importance of organizational goals has been pointed out several times, and cannot be over-emphasized (e.g., Murphy, 1987; Smith, 1976). Further, performance is not the consequence(s) of action, but rather the action itself. "Solutions," "statements," or "answers" produced as a result of unobservable cognitive "behavior," however, are defined as performance. In the job context, the fundamental distinction between measures of performance and measures of results is the degree to which "results" are a function of factors other than an individual's actions.

It is axiomatic in this model that performance is not one thing. A job, is a complex activity; and for any job, there are a number of major performance components (i.e., factors, constructs, dimensions) that are distinguishable in terms of their determinants and covariation patterns with other variables.

Determinants of Performance

Individual differences on a specific performance component are viewed as a function of three major determinants: (1) declarative knowledge, (2) procedural knowledge and skill, and (3) motivation (McCloy, 1990). Declarative knowledge (DK) is knowledge about facts and things (Anderson, 1985; Kanfer & Ackerman, 1989). Procedural knowledge and skill (PKS) are attained when declarative knowledge (knowing what to do) has been successfully combined with knowing how to do it (modified from Anderson, 1985; Kanfer & Ackerman, 1989). As a direct determinant of performance, motivation (M) is defined as the combined effect of three choice behaviors: (1) choice to expend effort; (2) choice of level of effort to expend; and (3) choice to persist in the expenditure of that level of effort. Of course, performance differences can also be produced by situational effects such as quality of equipment or differences in the degree

of external support across individuals. For purposes of selection and classification research, however, situational determinants such as these should be kept constant as much as possible.

The precise function relating DK, PKS, and M to performance is not known and may not be knowable. The important implication is that performance is directly determined only by some combination of these three elements, and for a specific performance component in a particular job, the things which facilitate or predict each of the three are most likely different. Further, the variables which control each of the three choices comprising motivation are probably also different (Kanfer & Ackerman, 1989). Consequently, personnel selection and classification efforts will be more effective to the extent that the determinants of different performance components are understood and are significantly under the control of abilities, personality, interests, or previous experience that can be measured at the time of selection or classification. The major features of this representation of the determinants of major performance components are illustrated in Figure 3.

$$PC_i = f(DK \times PKS \times M)$$

where $i = 1, 2, 3 \dots K$ performance components

DK = **Declarative Knowledge** (facts, principles, goals, self knowledge)

PKS = **Procedural Knowledge and Skill** (cognitive skill, psychomotor skill, physical skill, self management skill, interpersonal skill)

M = **Motivation** (choice to perform, level of effort, persistence of effort)

Figure 3. Determinants of Job Performance Components (PC)¹.

¹Individual differences, learning, and motivational manipulations can only influence performance by increasing declarative knowledge, procedural knowledge and skill, or the three choices.

Note. Adapted from J. P. Campbell, R. A. McCloy, S. H. Oppler, and C. E. Sager, 1992, in N. Schmitt and W. C. Borman (eds.). Frontiers in industrial/organizational psychology: Personnel selection, San Francisco: Jossey-Bass.

Latent Structure of Performance

The model includes a substantive specification for eight general performance components which are viewed as a reasonable starting point, or organizing framework, for considering performance measurement in virtually any job. Actual job analysis information would make the framework more specific to the jobs or positions being studied. The model invokes no general performance dimension. Rather it is hierarchical with eight performance components at the most general level. Across jobs, the eight dimensions have different patterns of sub-dimensions and their content varies

differentially. Further, any particular job might not include all eight dimensions. Below we describe each of these dimensions with regard to jobs in general. These definitions have been extracted from those provided in J. P. Campbell et al. (1992).

1. Job specific task proficiency. This dimension reflects the degree to which an individual can perform the core substantive or technical tasks that are central to his or her job. They are the job specific performance behaviors that distinguish the substantive content of one job from another. Doing word processing, designing computer architecture, maneuvering a tank through an urban environment, and sending and receiving radio messages are all examples of job-specific task content. Individual differences in how well such tasks are executed is the focus of this performance dimension. The question of how well individuals can do such tasks is meant to be independent of their level of motivation or the degree to which they contribute to effective group interaction. Technical or substantive task proficiency is also to be distinguished from supervisory or management tasks that involve interpersonal influence and the coordination of the work of others.

2. Non-job-specific task proficiency In virtually every organization, individuals are required to perform tasks or execute performance behaviors that are not specific to their particular job. For example, in research universities with Ph.D. programs, the faculty must "teach classes," "advise students," "make admission decisions," and "serve on committees." All faculty must do these things, in addition to doing research in their areas of expertise. In the Army, this factor is institutionalized as a set of common tasks (e.g., first aid, basic navigation) for which all soldiers are responsible.

3. Written and oral communication task proficiency. Many jobs in the work force require the individual to make formal oral or written presentations to audiences that may vary from small to large. For those jobs, the proficiency with which one can write or speak, independent of the correctness of the subject matter, is a critical dimension of performance.

4. Demonstrating effort. This dimension is meant to be a direct reflection of the consistency of an individual's effort day-by-day. It includes the degree to which the worker will expend extra effort when required, and the willingness to keep working under adverse conditions. It is a reflection of the degree to which individuals commit themselves to all job tasks, work at a high level of intensity, and keep working despite adverse conditions.

5. Maintaining personal discipline. This dimension is characterized by the degree to which negative behaviors, such as alcohol and substance abuse, law or rules infractions, and excessive absenteeism are avoided.

6. Facilitating peer and team performance. This dimension is defined as the degree to which an individual supports peers, helps them with job problems, and acts as a *de facto* trainer. It also encompasses how well an individual is committed to the goals of the work group and tries to facilitate group functioning by being a good model, keeping the group goal directed, and reinforcing participation by the other group

members. Obviously, if the individual works alone, this dimension will have little importance. However, in military jobs, high performance on this dimension is likely to represent a major contribution toward the goals of the organization.

7. Supervision. Proficiency in the supervisory dimension includes all the behaviors directed at influencing the performance of subordinates through face-to-face interpersonal interaction and influence. Supervisors set goals for subordinates, teach them more effective methods, model the appropriate behaviors, and reward or punish in appropriate ways. The distinction between this dimension and the previous one is a distinction between peer leadership and supervisory leadership. While modeling, goal setting, coaching, and motivating are elements in both dimensions, it is hypothesized that peer vs. supervisor leadership implies significantly different determinants.

8. Management/administration. This dimension is intended to include the major elements in management that are independent of direct supervision. It includes the performance behaviors directed at articulating goals for the unit or enterprise, organizing people and resources to work on them, monitoring progress, helping to solve problems or overcome crises that stand in the way of goal accomplishment, controlling expenditures, obtaining additional resources, and representing the unit in dealings with other units.

Summary. The eight dimensions described above are meant to be the highest order performance structures that could be useful. It is hypothesized that collapsing them into a smaller set would obscure important job information. Also, not all the dimensions are relevant for all jobs. What the model asserts is that the eight dimensions, or some subset of them, can describe the highest order latent variables for every job.

It should be emphasized again that the factors described above represent an organizing framework from which to begin investigating the components of performance in any particular job. Some dimensions may not be relevant for a particular job, and the nature of the sub-dimensions will vary across jobs or sets of jobs. Information from literature reviews, interviews with subject matter experts, and other job analysis data sources are needed to elaborate upon, verify, and modify the initial model.

Analysis of the JPM Project's Coverage of Performance

Coverage of Performance Dimensions. When viewed from the perspective of this model, important differences emerged across the Services in their approach to performance measurement. The Army attempted to measure all the important dimensions of performance that were identified through task-based and behavior-based job analysis. This roughly corresponded to dimensions 1, 2, 4, 5, and 6 above. In contrast, the other Services focused almost exclusively on job/occupation specific task proficiency (i.e., dimension 1). Ratings tapping other aspects of performance collected by the Air Force and Navy were generally excluded from validation analyses.

So just how much of the criterion space needs to be covered? Consider that there are many reasons why an organization might wish to measure job performance, and the preferred measurement strategy will be dependent upon the nature of those goals. With

regard to validation needs, the Services basically have a two-stage system that they must support in their research (Russell et al., 1992). The first stage, selection, is used to determine if an individual will be able to meet general performance requirements imposed by the organization (e.g., willingness to work hard and stick with the job). The second stage, classification (placement), is used to determine the jobs in which the individual is likely to perform most successfully.

The Army research was designed to validate not only the ASVAB, but other experimental cognitive and noncognitive selection and classification measures as well. Some of these measures were likely to be more useful for predicting some performance components than others and to have different importance for selection versus classification. To provide a reasonable test of these measures' predictive value, therefore, the Army focused on multiple components of performance. The other Services, however, were interested primarily in evaluating the predictive validity (both in terms of selection and classification) of ASVAB, and in some cases, other cognitive ability measures. Such measures seem most important for predicting technical task proficiency.

The way in which non-job-specific task proficiency was handled across the different Services appears to have been determined primarily by differences in force management aids. The Army has a clearly defined set of tasks which are required of all soldiers, regardless of Military Occupational Specialty (MOS) or occupational field. Thus, the Soldier's Manual of Common Tasks forms the basis for dimension 2 of the Campbell et al. performance model. Although the other Services may have job requirements that are force-wide, they are not clearly delineated in any documents of which we are aware. Thus, they tested no tasks that obviously fit into dimension 2 of our performance model. There were cases, however, in which tasks were identified as being common to an occupation. The Marine Corps studied several MOS within a single occupational field whereas the other Services generally tested only one job from a given occupational field. Thus, a set of tasks common to all infantry MOS (or all helicopter maintenance MOS) were identified, as were sets of tasks specific to each MOS in the relevant occupation (e.g., rifleman, mortarman).

Determinants of Performance. One can conceptualize the difference between maximal and typical performance measures as reflecting the degree to which the motivational determinant is allowed to influence scores on specific performance factors. Maximal performance measures essentially hold motivation constant whereas typical performance measures do not. Maximal performance measures are analogous to standardized testing conditions in which everyone is motivated to do as well as they can. Measures of typical performance are analogous to observations of actual job performance in real time which implies that motivation can play whatever role it typically does. Large differences in maximal task proficiency and typical task proficiency have been found (Sackett, Zedeck, & Fogli, 1988). The Services largely opted for the measurement of maximal task proficiency in the JPM project. To some extent, this was based on the conclusion that we do not know how to measure typical performance well enough in real time to justify doing so (Wigdor & Green, 1991). The Army attempted to capture typical performance using peer/supervisory ratings and a large array of

administrative indices of performance. Indeed, the ratings collected by all of the Services provided some indication of typical performance.

Because the goal of standardization is not entirely compatible with the goal of including motivation as a source of variance in measures of performance, assessment of typical performance is indeed difficult. The inclusion of typical performance information in a set of criterion measures, however, will permit a more accurate reflection of actual on-the-job performance.

IV. MEASUREMENT METHODS AND EVALUATION FACTORS

In the preceding chapter, we discussed the need to examine criterion measurement in the context of a guiding model of performance. The purpose of this chapter is to lay the remainder of the necessary groundwork for subsequent chapters which present our current understanding of the value of various criterion measurement methods. Specifically, in this chapter we will describe a framework for classifying criterion measurement methods and review the characteristics of criterion measures which can be used to evaluate their utility for validation research. In this chapter, we assume that we know what we want to measure. That is, the issues discussed herein relate primarily to "how to measure" rather than "what to measure."

Criterion Measurement Methods

It is difficult to classify various criterion measurement methods into distinct categories. For example, distinctions between work samples and relatively high fidelity simulations are fuzzy at best. Moreover, a "hands-on" work sample may be in the form of a paper-and-pencil test if the task being tested so warrants (e.g., a decoding task). Because one cannot necessarily equate a paper-and-pencil test with a simple assessment of knowledge rather than skill, then one cannot presume that hands-on and written tests are conceptually distinct entities. Furthermore, with rapidly advancing measurement technology, tasks that cannot be simulated in a paper-and-pencil format may, nevertheless, be abstracted to fit this measurement method with less loss of realism than traditionally expected. For example, advances such as performance-based multiple choice questions and critical incident-based problem scenarios can increase the fidelity of written "knowledge" tests to further narrow conceptual distinctions between this method and others (e.g., work samples).

Despite the fact that specific criterion measures may be difficult to classify into conceptually distinct categories, a comprehensible review of the criterion measurement literature requires the use of a classification scheme, no matter how rudimentary. Guion (1979a) defined a test as being a stimulus --> response mechanism designed for assessment purposes. He then proposed a taxonomy of measurement methods which was based on the manner in which performance is observed, recorded, and scored (i.e., the "response" part of the test mechanism). In contrast, we have chosen to organize our discussion of measurement methods around a classification scheme which is based loosely on the mode of test administration (i.e., the "stimulus" part of the test mechanism). Keep in mind, however, that both the "stimulus" and "response" characteristics of a test will contribute to (or detract from) the realism with which test scores will reflect actual job performance levels. Thus, both types of characteristics will be discussed in this report.

Our organizing scheme, then, is as follows:

- Direct Measurement
- Performance Tests
- Verbal Tests
- Job Performance Ratings
- Archival Records

Direct measurement of performance does not require simulation of task stimuli. Rather, a sample of actual performance on the job is recorded and scored. *Performance tests* (e.g., *work samples*, *simulations*) use actual work equipment or computerized, audio, and/or video aids to simulate task stimuli. *Verbal tests* depict task stimuli using words which may be supplemented with static pictures, drawings, and so forth. Although such measures are usually administered in a paper-and-pencil format, computerized and oral administration is becoming increasingly common. The fourth category, *job performance ratings*, comprises retrospective assessments of actual performance by judges who must rely on performance information stored in memory. Such ratings may be made by any knowledgeable individuals who have had the opportunity to witness relevant behaviors (e.g., peers, supervisors, self, clients). This method is distinguished from direct measurement by the fact that observations of performance were not done specifically for purposes of research-centered measurement. Finally, indices of performance which can be collected from an organization's *archival records* may incorporate one or more of the other measurement method formats (e.g., operational performance ratings). The distinguishing feature of this last category is that the performance information was collected and stored for reasons other than providing research criteria (i.e., there is no specially-developed measurement "stimulus").

Characteristics of High Quality Criteria

The quality of a given criterion measure is dependent upon a number of factors. Of particular importance are the following considerations:

- Relevance
- Comprehensiveness
- Susceptibility to contamination
- Reliability
- Discriminability
- Practicality

Any criterion score or set of criterion scores, whether it is based on a single measurement method or on composites based on multiple methods, can and should be evaluated on the basis of these factors. Each of the factors is discussed briefly below. The first three factors (i.e., relevance, comprehensiveness, and susceptibility to contamination) embody questions of content and construct validity.

Relevance. Relevant criteria measure behaviors and/or traits which are related to the accomplishment of organizational goals, and which the organization expects job

incumbents to be able to perform. To maximize the likelihood that a criterion measure is relevant, it should be developed on the basis of a carefully defined, job-analysis-based, job content domain (Guion, 1979b).

The question of relevance applies both to the content of the measure (i.e., which tasks are covered) as well as to the fidelity of the measurement operation (i.e., how performance on the criterion content is measured). The three determinants of performance discussed in Chapter III, DK, PKS, and M are relevant to the question of measurement fidelity. A criterion measure which allows all three of these performance determinants to contribute to an individual's score exhibits higher fidelity (i.e., captures actual job performance more accurately) than a measure which holds one or more of these determinants constant. For example, direct measurement, in which performance on the job is recorded unobtrusively for some period of time, allows all three performance determinants to operate fully. A work sample which requires task performance outside of the normal work situation holds motivation constant, and is therefore a lower fidelity measurement method.

Measurement fidelity is a function of several test characteristics, including item type, mode of item delivery, context of item delivery, mode of examinee response, and scoring procedures. All of these factors contribute to fidelity, although all too often, the role of the scoring process in this regard is overlooked. For example, a work sample test which reproduces job demands very realistically from all outward appearances, but which does not incorporate a scoring strategy that accurately reflects the determinants that influence performance has significantly limited fidelity. For the rating method, the critical consideration is the fidelity with which judges' ratings are based on unbiased observations of representative performance behavior.

Comprehensiveness. The extent to which a given criterion measure, or set of criterion measures, assesses all significant aspects of performance is the basis of this evaluation factor. In the language of our general performance model described in the previous chapter, a set of criterion measures is comprehensive to the extent that each relevant dimension of performance is represented by a criterion score. A criterion score is comprehensive to the extent that it is based on all important aspects of performance within the dimension or sub-dimension it is intended to reflect.

Susceptibility to contamination. A significant threat to the validity of criterion scores is the extent to which score variance is attributable to job irrelevant determinants (i.e., anything other than DK, PKS, or M) that cannot be controlled for statistically. For example, promotion rate, an administrative index of performance, may be influenced by a host of factors that are not under the job incumbent's control (e.g., organizational politics, lack of opportunity). Race and sex bias in measurement is usually a primary concern for researchers. Contaminants may be introduced by the measurement method itself. Verbal measures may require vocabulary knowledge which is not required for successful task performance, and raters may provide performance evaluations that are biased by factors such as knowledge of performance on predictor measures.

Reliability, as a characteristic of a specific measurement method, refers to the consistency or variation in scores that would be obtained if the method were used to measure the same individual a number of times, and the individual's true score on the latent variable of interest did not change. The most direct indicator of measurement consistency or precision is the standard error of measurement or standard deviation of one person's score when measured many times. Given a constant true score, the degree of unsystematic variation in the observed score reflects reliability, or precision of measurement. Sources of systematic variation, still given a constant true score, are what is meant by contamination (e.g., variation due to race, gender, or appearance that has nothing to do with the true score). Unsystematic variation could result from changes in the measurement operations (e.g., a slightly different sample of items, a different hands-on test examiner), from unsystematic influences attributable to the person (e.g., variations in fatigue levels), or to the environment (e.g., lighting fluctuations).

Measuring individuals many times with the same method, while holding true score constant, is usually experimentally difficult. Various reliability models finesse the problem by making a few simplifying assumptions (e.g., measurement errors are a random normal deviate) and then deriving various indices by which the precision or consistency of measurement can be estimated (e.g., the standard error of measurement, or the reliability coefficient). For example, *test-retest reliability* estimates are applicable when the performance being assessed is not likely to change substantially between two test administration sessions giving the same measure twice, and will not change the individual's true score. Thus, the correlation between scores collected on the same test at two different points of time would be an appropriate assessment of the reliability coefficient. *Internal consistency* reliability estimates (e.g., coefficient alphas) are useful when a test is intended to measure a unidimensional aspect of performance because these estimates treat performance variability within the test as error. Because performance, even at the task level, is often multidimensional, internal consistency reliability is somewhat less applicable to criterion measures than to single construct-based predictor measures. *Estimates of equivalence* indicate the extent to which two or more measures of the same attribute produce equivalent scores and are suitable when parallel forms of a test are developed. Conceptually similar indicators, *coefficients of agreement*, estimate the extent to which different judges yield equivalent ratings of performance.

The types of reliability estimates listed above are founded on classical test theory. Each type of reliability estimate treats measurement error as a unidimensional contributor to score variance. Generalizability theory, however, has provided researchers with a more sophisticated way to examine measurement error in that it allows them to postulate different sources of error and to simultaneously examine the impact of multiple potential sources of error using analysis of variance strategies (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Within the limitations of the data collection design, a researcher may use generalizability theory to examine the impact of using different scorers, testing different sets of tasks, testing in different locations or under different conditions, and so forth. Unfortunately, data collection designs are not always conducive to such analyses because they are not sufficiently crossed (e.g., scorers may be nested within test locations and tasks). When the right data are available, however, this analysis strategy is an informative way to examine measurement error. General discussions regarding the

application of generalizability analyses to JPM project data are provided by Shavelson (1991) and Kraiger (1989).

Discriminability. A discriminating criterion measure distinguishes among good and poor performers. In validation research, the score distribution has to have enough variance to allow any covariance with predictor measures to manifest itself. This standard will be difficult to meet to the extent that there is relatively little real variation in performance. Such a situation would arise, for example, in an organization that could quickly remediate or remove below average performers. The sheer size of military occupations, however, helps to ensure that a reasonable amount of performance variability is there to be captured.

In the context of task-based measures, discriminability can be fine-tuned by increasing or decreasing the difficulty of the tasks which are selected for measurement and/or increasing or decreasing the difficulty of the tests which are based on those tasks. Efforts to ensure discriminability in ratings usually center on the development of suitable rating scale anchors and the use of rater training programs (e.g., Pulakos, 1986).

Practicality. The human resource and financial costs of developing and administering criterion measures are the primary determinants of a measure's "practicality." The Joint-Service JPM effort was an admirable effort to expand the usual bounds of practicality, but such a luxurious environment for the criterion measurement researcher is a very rare circumstance. Future military performance measurement efforts, like those of most other organizations, will probably be greatly constrained by the costs associated with high quality research. Therefore, a consideration of the practical side of criterion measure alternatives is clearly warranted.

Format for Criterion Measure Review

In the following chapters, specific criterion measurement methods will be reviewed and discussed. The specific types of measures, within each major measurement method, that will be discussed in the remainder of this report are summarized in Figure 4. Note that "direct measurement" is not included as a separate topic of discussion. Typically, only those jobs which permit machine recording of important behaviors (e.g., telephone operators, airline reservation clerks) or simple tallying of observable behaviors by unobtrusive human observers could use of this measurement strategy. Because obtaining quality criterion scores using this measurement method is not likely to be feasible for most jobs, and because there is relatively little literature to draw upon for a review, we have chosen not to provide a separate chapter for direct measurement in this report.

Our discussion of measurement methods will center around the evaluation factors presented above. In addition, we will specifically focus on the criterion measures developed in the Joint-Service JPM project. This focus will allow us to consider the following research questions: (1) What do the JPM projects tell us about the validity of ASVAB and the experimental predictor measures developed by the Army and Navy? (2) What do the JPM projects fail to tell us about the validity of ASVAB and the

experimental predictor measure? (3) What criterion measures show the greatest potential for future selection and classification research in the military?

PERFORMANCE	VERBAL
Work Samples	Traditional multiple-choice
Simulations	Performance-based multiple-choice
<ul style="list-style-type: none"> • Computer/Visual/Audio Aids • Assessment Center Exercises 	Written Essay
	Oral Interview
	Accomplishment records

RATINGS	ARCHIVAL
Task-based	Training grades
Dimensional/Job-specific	Supervisor ratings
Global/Force-wide	Promotion rate
Supervisor	Rewards
Peer	Disciplinary actions
Self	Production Indices
	Turnover

Figure 4. Summary of Measurement Methods.

V. PERFORMANCE TESTS

In our measurement method classification scheme, performance tests are characterized by the use of equipment and/or audio and visual aids to simulate task stimuli. Most of these types of tests require examinees to physically interact with the test stimuli, as in the case of a hands-on work sample test.

Historically, performance tests have been used more commonly as predictor rather than criterion measures, with a major exception being testing for professional certification and licensure programs. Despite their infrequent use for this purpose, however, high fidelity simulations and/or work sample tests have been viewed as particularly desirable criterion measurement methods because they call for the application of job knowledge and demonstration of job skills (e.g., Asher & Sciarrino, 1974). They do so by eliciting behaviors that are equivalent, or nearly equivalent, to those required in the job setting. Such measures also benefit from wide acceptance by decision-makers due to their face validity. Indeed, the *face validity* of hands-on testing is one of the primary reasons work samples were identified as the benchmark testing method in the Joint-Service JPM research program.

In this chapter, we will discuss two major types of performance tests: work samples and simulations. As with our primary test method classification scheme, the distinctions between these two types of performance tests are rather artificial. Nonetheless, the distinctions are useful to make for purposes of discussion. Thus, we define work samples as task-based hands-on tests which use actual work equipment and job aids as task stimuli. Work samples are exemplified by the hands-on tests constructed by the Services for the JPM research. We include, as well, the performance-based interviews constructed by the Air Force in this same research effort. In contrast, simulations are either task-based or knowledge, skill, and ability (KSA)-based tests which use props, rather than actual work equipment, to recreate work-related situations. We will discuss the use of tests which use computer, audio, and/or video aids and assessment center-type exercises as primary examples of simulation measures.

Within each type of performance test (i.e., work samples and simulations), we will review test development, administration, and scoring strategies. We will also discuss the adequacy of these types of measures using the evaluation factors outlined in Chapter IV.

Work Samples

Content Sampling

By its nature, performance testing is almost invariably task-based, at least on the surface. That is, work samples and simulations are based on task requirements. The scoring system, however, may be trait-based as is the typical case in assessment center testing.

Guion (1979c) outlined a methodology for selecting content for testing. An adaptation of that methodology includes the following four steps:

- Step 1. Use job analysis to define the *job content universe* which comprehensively describes job tasks.
- Step 2. Select tasks from the job content universe which are to be considered for testing. Criteria for inclusion may include frequency of performance, difficulty, type (e.g., technical, interpersonal), and so forth. This constitutes the *job content domain*.
- Step 3. Define the *test content universe* by eliminating tasks which cannot be adequately measured using available measurement methods.
- Step 4. Identify the *test content domain* by selecting a sample of tasks for testing from the test content universe. Task selection may be based on content representativeness, time available for testing, and/or other considerations.

For selecting tasks for JPM hands-on testing, each of the Services used sampling procedures which can be characterized using this basic model. Their efforts to ensure content representativeness were aided by the use of task clusters which organized tasks by either co-performance or content similarity. Details of each of the task sampling strategies used by the Services are summarized below.

Air Force. The Air Force defined the job content universe for each sampled AFS using task lists generated through its ongoing occupational analysis program (Lipscomb & Dickinson, 1988). The job content domain was constructed by first reviewing and "cleaning up" the universe of tasks. This included the elimination of redundant or outdated tasks. After this clean-up, tasks which were taught in formal training and/or were performed by a significant percentage of first term incumbents (e.g., 30% or more) were identified for inclusion in the job content domain. These tasks were also organized by co-performance and task difficulty using data retrieved from the occupational analysis data base.

The Air Force did not identify a test content universe, per se. Rather, researchers constructed a test content domain using an iterative process that began with the selection of a set of tasks using a stratified random sampling plan in which tasks were stratified by function and difficulty level (Lipscomb & Dickinson, 1988). If a sampled task was not suitable for performance testing or if SMEs considered it otherwise inappropriate for testing, it was dropped and replaced with another randomly selected task.

The task sampling process was generally performed for two levels of job content (Lipscomb & Dickinson, 1988). To understand these levels, it is important to note that most AFS are appropriately regarded as groups of similar jobs rather than a single job. This is due to the fact that few tasks are performed by all AFS incumbents, and most tasks are performed by one or more subgroups (or job types) within each AFS. In light

of this diversity, the more populated job types within each sampled AFS were selected for performance measurement. For example, only Jet Engine Mechanics who worked on J-57, J-79, or FT-33 engines were tested. Thus, tasks selected for testing represented a combination of AFS-specific and job type-specific job requirements.

Army. Army researchers identified the job content universe for each MOS by consolidating information from four major sources - the Manual of Common Soldiering Skills, MOS-specific Soldier's Manuals, the Army Occupational Survey Program (AOSP) data base, and subject matter experts (SMEs) (C. H. Campbell et al., 1990). These sources provided information on technical tasks only. The job content domain was identified with the assistance of three SMEs. In this phase, tasks which were outdated, low-frequency or team-based were eliminated from consideration. As with the Air Force's approach, the job content domain and test content universe were synonymous.

Fifteen SMEs (1) ranked tasks by importance in multiple scenarios (e.g., garrison, combat), (2) sorted tasks into clusters, and (3) provided an estimate of task performance variability (C. H. Campbell et al., 1990). Using these data along with AOSP frequency data, job analysts identified the written test content domain. The hands-on test content domain included those tasks in the written test content domain which were particularly suited for hands-on testing and which adequately represented the task clusters. SMEs had the opportunity to review and approve the final test content domain specifications. Similar to the Air Force approach, test content domains were identified at two levels. In this case, however, one level was Army-wide and the other level was MOS-specific.

The longitudinal validation component of Project A/Career Force included the development and administration of hands-on tests and role-play simulations for second term soldiers. The procedures used to select technical tasks for testing were essentially the same as those used for first term incumbents (J. P. Campbell & Zook, 1990b). Because the Army's AOSP data base does not incorporate supervisory-related tasks, however, additional job analysis effort was required to identify these requirements. A task-based analysis used previously-developed Army instruments to comprise the job content universe (Leader Requirements Survey - Steinberg & Leaman, 1987; Supervisory Responsibility Questionnaire - White, Gast, & Rumsey, 1986). The tasks on these surveys were consolidated, and the most important tasks (n=53) were added to the job content domain for each MOS. They were then incorporated into the task selection process as a third level of measurement (i.e., common technical, MOS-specific technical, and supervisory).

Navy. The Navy used two different sets of procedures for selecting tasks for testing in its JPM effort, one for the Radioman rating and another for remaining ratings. For both sets of procedures, a rating's job content universe was generated by reviewing task lists from the Navy's occupational analysis data base and training materials, and supplementing these with SME input (Laabs & Baker, 1989). To create the job content domain, all tasks which were not rating-specific and technical were eliminated. At this point, as well, additional clean-up of the task lists took place (e.g., consolidating similar tasks), and the remaining tasks were clustered into functional areas using

multidimensional scaling (Laabs, Berry, Vineberg, & Zimmerman, 1987). At this point in the process, the two sets of sampling procedures used by the Navy diverge.

For the Radioman rating, the Navy essentially equated the job content domain with the test content universe, holding issues regarding suitability for hands-on testing to the end of the process (Laabs & Baker, 1989). This same approach was used by the Army and the Air Force. A mail-out survey was prepared and administered to generate task characteristic data for use in the task sampling process. Specifically, incumbents and supervisors provided data on frequency of task performance, task criticality, and task difficulty. To construct the test content domain, tasks were selected in successive-hurdles fashion using this information and task cluster membership. A secondary list was made to provide replacements for any tasks that were considered unsuitable for hands-on testing. These decisions were made with SME input.

For the other Navy ratings, the test content universe was created by eliminating tasks that could not be hands-on tested in a standardized fashion (Laabs et al., 1987). In workshop settings, SMEs identified the most important tasks and grouped these tasks into two sets of categories: functional and behavioral. Researchers then identified tasks for testing using several selection algorithms in conjunction with each other. Specifically, a subset of tasks was selected using the behavioral categories as the major stratification criterion, another subset was selected using the functional categories as the major stratification criterion, and a third subset was selected in which a functional and behavioral category weighting scheme was used. In all three selection algorithms, task importance and task difficulty were significant determining factors for task selection.

Marine Corps. The Marine Corps used training doctrine documents (particularly the Individual Training Standards (ITS)) and SME input as the basis for establishing the job content universe for the jobs they studied (Crafts et al., 1991). The job content domain was essentially equivalent to the job content universe. In the construction of the test content universe, tasks were dropped for a variety of reasons. For example in the mechanical maintenance MOS, tasks performed very differently on different pieces of equipment (e.g., different types of trucks) were dropped because mechanics do not usually have experience working on many different types of vehicles. Other reasons included infeasible equipment requirements and potential for injury. At this point, tasks were also identified as being required of all Marines in the occupational field or for a given MOS only (i.e., core or MOS-specific). In addition, tasks were organized into functional duty categories. For the mechanical maintenance MOS, which had many more tasks than could be tested, SMEs identified the most important cells (each of which contained some number of tasks) to include in the test content universe.

The Marine Corps added a unique feature to their job analysis procedures. For each task, they identified "behavioral elements" which underlie task performance (Crafts et al., 1991; Felker et al., 1988). Behavioral elements are intended to capture the essential behaviors required to perform a task which may be common to other tasks (e.g., check for dirt, rust, or damage; conduct function check). Theoretically, performance on tasks having many behavioral elements in common could be predicted by the same abilities and would exhibit similar performance levels.

The test content domain was identified by randomly selecting tasks for testing from each cell in the task-by-element matrix. Each time a task was selected, the job or behavioral elements required by the task were noted. Tasks which were overly redundant with the elements already covered were withdrawn from the pool. Thus, the elements were used in the task selection procedure in an effort to reduce redundancy in the sample of tasks with regard to their requisite knowledges, skills, and abilities. Selection of core and MOS-specific tasks was conducted independently.

The Marine Corps actually selected two sets of tasks to represent the core infantry domain of tasks (Felker et al., 1988). The resulting tasks then formed the basis of comparable forms (sets) of the core hands-on tests. This would allow an examination of the reliability of the task selection procedures and resulting performance measures.

Summary. Despite many procedural differences, the fundamental approach to task selection was similar across the Services' JPM efforts. With the exception of the Army's second term supervisory tasks, the job content domain was restricted to individually-performed technical tasks. With the exception of the Army, all tasks were job/occupation-specific. With respect to this difference, it appears that the other Services do not have an institutionally defined set of tasks that were performed by all enlisted incumbents (e.g., maintain an M16 rifle) which would permit them to have Service-wide measurement at the task level.

Selection for the criterion measure content domain was based on one or more indicators of importance, difficulty, and/or frequency provided by SMEs, and a goal of content representativeness. Methods for obtaining these indices and for establishing content representation varied widely, but there is little reason to believe that the methods produced appreciably different test domains. Regardless of whether they were dropped before, during, or after initial task selection efforts, all of the Services had to eliminate tasks from performance testing eligibility. Reasons included (a) potential danger to personnel or equipment, (b) unreasonable time requirements, (c) infeasible to test in a realistic manner, and (d) inaccessibility of equipment. Finally, all of the Services involved SMEs in the task selection process both in terms of providing judgments of importance, frequency, and so forth, and also in final approval of the task domain lists.

The Marine Corps took a labor-intensive extra step to minimize redundancy of job information with its use of behavioral elements in the task selection process. To the extent that this step appreciably reduced redundancy of performance information without undue loss of measurement reliability, it may be worth including in future task selection exercises. Analysis of Marine Corps data to examine the correspondence between tasks sharing behavioral elements might help to determine the payoff of this procedure.

Both the Air Force and Marine Corps incorporated a random sampling component into their task selection strategies. This was at least in part a response to the NAS oversight committee's guidance on task selection (Wigdor & Green, 1991). In the NAS view, JPM project job performance scores should reflect the amount of the job that an examinee has mastered so that the scores would more adequately support the Services' enlistment standard setting activities. As an example, an absolute competency

score of 70 percent correct on a set of hands-on task tests would indicate that an examinee had mastered to some standard 70 percent of the tasks that comprise his or her job. There are several assumptions that must be met, however, for this statistically-based score interpretation to be meaningful. First, there must be a "standard" for each task that defines whether or not it has been mastered. Another critical assumption was that random selection from the job-content universe is possible so that statistical generalizations from sampled tasks to the whole job can be made.

As the Air Force and Marine Corps approaches to task selection illustrate, random selection necessarily has to be performed in conjunction with a great deal of prior adjustment to the job content universe (e.g., eliminating tasks that cannot be tested in a hands-on mode, eliminating trivial tasks). The interpretability of statistical generalizations made on the basis of such a contrived job domain strikes us as being no more useful than rationally-based score interpretation. In other words, the random sampling component probably does not hurt, but it is doubtful that it helps very much either with regard to the meaningfulness of the resulting performance assessment. Wigdor and Green (1991) propose another alternative to purposive sampling in an effort to preserve some capability to generalize to the job content universe with a specifiable degree of error. They suggest selecting many random samples and allowing SMEs to eliminate samples that do not strike them as reasonable. The meaningfulness of the resulting statistical generalizations that might be made, however, remains debatable. The general issue of standard setting is discussed at greater length in the Task 5 Roadmap report (McCloy, 1992).

Construction

All of the Services used active duty SMEs to assist in the delineation of steps involved in performance of the tasks to be tested in a work sample mode. In each case, tasks were broken down into a series of observable steps, and critical performance standards were noted. The Marine Corps was substantially aided in this effort by its Individual Training Standards (ITS) which specify the steps required for task performance and the minimum competency requirements for each task. The Army's training manuals provided similar support. For example, these documents specify time limit requirements (e.g., put on protective mask within fifteen seconds) and indicate whether task steps have to be performed in a particular order for satisfactory performance. In many cases, it was not possible to test a task in its entirety due to practical constraints such as time or safety considerations. In some of these cases, essential testable elements of the task were identified and comprised the work sample test. In other cases, the examinee would be asked to explain what should be done rather than actually performing the time-consuming portion of the task.

In addition to specifying the steps to be scored within each task test, the hands-on tests also specified a set of standardized test conditions and test set-up instructions. A sample hands-on test form used by the Air Force is provided in Figure 5. Test forms for the other Services are essentially similar to this example.

Phase I J-79, J-57, TF-33
Shop and Flightline

Hands-On Task 347

Objective: To evaluate the incumbent's ability to install starters.

Estimated Time: 25 M **Start:** **Finish:** **Time Req:**

Time Limit: 35M **#Times Performed:** **Last Performed:**

Tools and Equipment: Consolidated Tool Kit, 0- to 150-inch-pound Torque Wrench, 10- to 300-inch-pound Torque Wrench, Lubricant.

Appropriated T.O.:

J-79 (Fighter):	1F-4E-10
J-57 (Tanker):	1C-135(K)A-2-4JG-6
TF-33 (Cargo):	1C-141A-2-4JG-5 or 1C-141B-10
General Torquing	2-1-111 or 1-1A-8 or specific engine torquing T.O.:
	J-79: 2J-J79-86-7WP00100
	J-57: 1C-135(K)A-2-4JG-1
	TF-33: 1C-141B-10

Background Information: There are some common steps for all three engines, but each engine has some unique steps. The evaluation will be made on the common steps except when indicated. Difference include:

1. J-57 has two cannon plugs.
J-79 and TF-33 (P7) have one cannon plug.
2. J-57 and TF-33 have one nut on the V-clamp.
J-79 has two nuts on the V-clamp.

Two-person task when actually putting the starter in place. This is the only task for which the incumbent will be required to actually get the technical order from the shelf.

Engine Configuration: The starter adapter pad must be on the engine. The starter is off the engine.

Instructions: Administer in the shop.
The incumbent MUST use the T.O.
Compare the incumbent's response to the correct answer for the appropriate engine.

Figure 5. Sample Air Force Hands-On Test.

SAY TO THE INCUMBENT

GET THE T.O. USED TO INSTALL A STARTER AND THE T.O. FOR GENERAL TORQUING PROCEDURES, THEN INSTALL THE STARTER USING THE APPROPRIATE PROCEDURES FROM BOTH T.O.S. FOLLOW GENERAL MAINTENANCE PROCEDURES AT ALL TIMES. TELL ME IF YOU PLAN TO DEVIATE FROM THE T.O. YOU MAY NOT ASK ANYONE TO HELP YOU FIND THE CORRECT T.O.

Performed or Answered Correctly	Yes	No
Did the incumbent:		
1. Obtain the appropriate T.O. for the starter installation and the torquing procedures within 10 minutes?	—	—
2. Hang the clamp per the specific T.O.?	—	—
3. Lubricate the spline?	—	—
4. Ensure that the starter was not left in an unsupported position (hung by the shaft) at any time?	—	—
5. Index (position) the starter per the appropriate T.O. J-79: Breech at 8 o'clock position J-57: Breech at 3 o'clock position TF-33: Drain plug at 6 o'clock position	—	—
6. Properly seat the V-Band Clamp?	—	—
7. Torque the V-Band Clamp per the appropriate T.O. J-79 Airsearch: 110 to 130 inch-pounds J-79 Sunstrand: 65 inch-pounds J-57: 65 to 70 inch-pounds TF-33: 60 to 70 inch-pounds	—	—
8. Install the locking device on the V-Band Clamp per the appropriate T.O.?	—	—
9. Connect the applicable electrical connector (cannon plug)? (Must not connect the tachometer generator plug on the J-57).	—	—

Figure 5. Sample Air Force Hands-On Test (Continued).

10. Use the correct tools and materials? _____

STOP TIME: _____

OVERALL PERFORMANCE

- 5 Far exceeded the acceptable level of proficiency
- 4 Somewhat exceeded the acceptable level of proficiency
- 3 Met the acceptable level of proficiency
- 2 Somewhat below the acceptable level of proficiency
- 1 Far below the acceptable level of proficiency

Figure 5. Sample Air Force Hands-On Test (Continued).

Note. From The methodology of walk-through performance testing by J. W. Hedge, 1987, in J. W. Hedge and M. S. Lipscomb (Eds.) Walk-through performance testing: An innovative approach to work sample testing, AFHRL-TP-87-8, Brooks AFB, TX: Air Force Human Resources Laboratory.

The notion of "scoreable" steps is not restricted to direct observation of task performance (i.e., process scoring). Some tasks result in a product which should be scored, either instead of or in addition to the process, to provide a comprehensive assessment of task performance. For example, completion of a range card by a cannon crewman or a typing test by an administrative clerk can only be evaluated when the product of the examinee's efforts is examined.

The Air Force combined work sample and interviewing testing strategies into a "walk through performance testing" (WTPT) method. As with the hands-on tests, the interview-based testing format also divided tasks into steps that were scoreable in a go/no-go fashion. Instead of actually performing the task, however, incumbents were asked to describe how they would perform the task in a show and tell fashion using actual equipment as visual aids. A subset of the tasks tested in the interview format were

also tested in the hands-on format to allow for a comparison of these two measurement methods.

One of the difficulties inherent in the development of work sample tests, including the interview-based variation described above, is differences in the way a task might be accomplished. To some extent, this problem is alleviated in the military because there are usually documents that provide doctrinally-established procedures for task performance. When the task is performed somewhat differently in a given unit or location, the test may be written to reflect doctrine. In other cases, however, procedural differences may be necessary to reflect different types of equipment or unusual environments (e.g., frigid weather). In these cases, minor changes will have to be made to the affected test's scoring system.

The initial development of a hands-on test is relatively simple provided a thorough task analysis is available which decomposes each task into discrete steps. Pilot testing is a critical phase, however, to the construction process. Even a very small-scale tryout of measures will identify problems such as (a) unanticipated equipment availability or unreliability problems, (b) performance steps which cannot be reliably observed, and (c) ambiguous instructions for scorers and/or examinees. The ability to observe task steps is not a trivial problem for some tasks. The problem is particularly severe for tasks which are performed in small spaces or which require close observation that interferes with the examinee's performance of the task.

Scorer Qualifications and Training

Even when hands-on test scoring is restricted to a simple checklist format, it is important that test administrators/scorers be knowledgeable about the tasks being tested. This is necessary because few tasks are so simplistic that they can be administered and scored with no technical knowledge relating to task performance. Tests which require the scorer to provide an overall proficiency rating, of course, also require the scorer to be intimately familiar with task performance requirements.

The Marine Corps hired retired senior NCOs to serve as hands-on test scorers. The Air Force used retired senior NCOs for two AFS and active duty senior NCOs for the other six AFS to administer and score WTPT performance. The Navy used active duty NCOs as scorers for the Radioman rating, but used retired NCOs for the remaining jobs. The Army used active duty NCOs to score hands-on tests for all jobs. All scorers were experienced in the MOS (or a related MOS) that they were testing. For the most part, the Services were reluctant to use active duty scorers for fear that the NCOs would be inclined to correct problems in task performance during the course of testing and because researchers would not be able to reject scorers with poor observation or scoring skills (e.g., Wigdor & Green, 1991). In other words, it was hypothesized that active duty scorers would produce less reliable and less valid performance scores. Although no direct comparisons between active duty and civilian scorers have been described, there are no obvious differences in hands-on data quality across jobs and services which correspond to the two different types of scorers.

Across all Services, scorer training was managed by civilian personnel who had participated in test development. All scorers were provided with instruction and practice

on test administration, observation, and scoring skills. Generally, training included an overview of the research project, an introduction to the work sample testing format, and a review of testing procedures (e.g., rotation of examinees through test stations). There was extensive use of role-playing in which scorers played either a scorer or an examinee to simulate testing and issues that might arise during actual testing (e.g., the examinee has no idea what to do next). The Air Force constructed videotapes to illustrate correct and incorrect examinee performance on the hands-on tests (Hedge, Dickinson, & Bierstedt, 1988). The videotapes were used to provide scoring practice and rater training.

Duration of scorer training varied considerably across the Services. The Army required one to two days, the Air Force and Marine Corps required one to two weeks. One reason for the variation is that the Army trained each scorer to staff only one test station whereas the Marine Corps trained each scorer to staff multiple test stations (each test station accommodates one or more task tests). The Air Force reduced the duration of scorer training from two weeks to one week as training procedures were refined throughout the course of JPM testing (Hedge & Teachout, 1992). Air Force scorers were trained to administer all task tests.

Test Administration

For each job, task tests were divided into groups so that there would be an even number of test stations, each requiring approximately the same amount of administration time. To the extent possible, tasks which are co-performed on the job were tested together. In the Air Force model, one scorer administered all tasks to an examinee (Laue, Bentley, Bierstadt, & Molina, 1992). In the Army model, each scorer staffed a single station, and examinees rotated among the test stations as they completed each station and were not required to complete the tests in any particular order. In the Marine Corps model, scorers periodically switched stations, and examinees rotated in unison according to a pre-established plan. All of the Services collected "shadow scoring" data. Shadow score data are collected by having a second scorer independently evaluate the examinee's performance simultaneously with the primary scorer. The Marine Corps included shadow scoring of selected examinees as a routine part of the test administration process. The Army conducted shadow scoring only for selected examinees in two MOS at two data collection sites.

A civilian hands-on test manager monitored the performance of the work sample scorers. In addition to on-the-job coaching, the Marine Corps used an on-site data entry system as a scorer feedback tool (Crafts et al., 1991; Felker et al., 1988). That is, hands-on data were key-entered soon after they were collected. Problems interpreting the scorers marks were identified and discussed with the scorer immediately. Each scorer's ratings could also be compared with the ratings of other scorers to identify deviations from the norm.

Scoring

All of the Services used a checklist (go/no-go) strategy for scoring both the process and products associated with hands-on performance. This greatly minimized subjectivity of the rating process. The Air Force, Army, and Navy (at least for the Radioman rating)

also included overall proficiency ratings for each task tested, but these ratings have not been included in validation analyses that have been reported. Most of the Services timed performance of all task tests but it is not clear how these data were used.

For the most part, the Services scored the work sample tests based on the proportion of steps performed correctly, and task scores were summed to yield a compensatory total performance score (Wigdor & Green, 1991). Specific scoring details for each of the Services were spotty or non-existent (J. P. Campbell, 1988; Hedge & Teachout, 1992). The Air Force weighted task steps by importance and used these importance weights for scoring purposes. The Marine Corps weighted tasks prior to combining task scores, at least for the Infantry MOS (Mayberry, 1988). Each of the Services' chosen scoring strategies, as well as many alternatives could be examined to determine the effects of various scoring options on validation results and the construct validity of the scores. The effects of weighting tasks or task steps, adjusting task scores for differences in number of constituent steps, adjusting performance scores for test site and/or job subtype differences, and adjusting task scores for differences in examinee experience are examples of scoring issues that could have an appreciable impact on the quality of the resulting criterion scores.

Evaluation

Relevance. The extreme care taken by the Services to identify job content universes leads us to feel confident that the tasks selected for testing were relevant to the jobs being examined. Of course, it is unlikely that researchers would select a task for work sample testing that is not actually performed by some job incumbents. Threats to relevance, however, come from steps taken to make a task testable and from having too many examinees inexperienced in the task. With regard to the first problem, a scenario for performing a task in a test situation may be so artificial and simplistic as to compromise the relevance of the test to the examinee's job.

The second threat to relevance is particularly severe for the Services which have thousands of incumbents in a given job. Under these circumstances, there may be many tasks that are not performed by all incumbents but which are very important for most incumbents. The end result is a test domain that has high relevance for job incumbents as a whole, but which includes one or more tasks that a given examinee does not perform on his or her particular job. In anticipation of this problem, most JPM examinees were asked to indicate their level of experience with each task on which they were tested. For example, Army examinees were asked two questions: (1) How often have you performed this task?, and (2) When was the last time you performed this task? Thus, analyses regarding the magnitude of this problem and its potential effect on validation results can be examined. Although analyses of these data have gone largely unreported, the Air Force found correlations ranging from .11 to .43 between work sample scores and task experience, suggesting a nontrivial influence of task experience on test performance (Hedge & Teachout, 1992).

Comprehensiveness. Administrative time and logistical complexities required by work sample testing restricts coverage of the intended performance domain. Thus, although one can usually measure a given task comprehensively, relatively few tasks can be tested. Each Service devoted four hours or more to work sample testing. Tasks were

carefully selected for testing to maximize coverage of the test content universe. The fact is, however, that for most jobs studied, work samples covered relatively few job tasks. As indicated above, the problem is exacerbated by the probability that a given examinee does not perform all of the tested tasks on the job, making coverage of his or her job even more limited.

The Air Force used the show and tell interview format as a strategy for expanding the content coverage of performance testing. Specifically, the oral performance interview strategy allowed the Air Force to test tasks that could not be feasibly tested using traditional work sample testing. Correlations between performance scores based on tasks tested both in a hands-on and interview format ranged from .34 (for Precision measurement equipment laboratory specialist) to .94 (for Personnel specialist), with the median correlation being .68 (Hedge & Teachout, 1992). The median correlation between the scores based on all tested tasks (i.e., not just tasks in common to both methods) was only slightly lower at .66. It is reasonable to believe that the nature of the task will be a major determinant of the correspondence between scores produced by these two measurement methods, and that this may explain the large differences in correspondence across jobs. In any case, interview-based performance testing appears to be a useful complement to traditional hands-on testing. However, it does not necessarily permit testing of more tasks, just more tasks that are critical.

Susceptibility to contamination. Two major sources of potential contamination for work sample tests are testing conditions and the involvement of the scorer. If all testing takes place at a single site, the set-up of equipment and other task stimuli (e.g., visual targets, barricades) is held constant. Outdoor testing, however, is at the mercy of the elements - rain, cold, snow, wind, and so forth. To the extent that one cannot outwait the weather, the potential impact of the weather on performance scores must be accepted. Furthermore, most of the JPM testing required set-up of work sample tests at many different locations. These locations differed with regard to physical characteristics, equipment, and in many cases, scorers. Although all efforts were made to standardize the set-up and administration of each test across sites, differences were bound to introduce some contamination into scores. R. G. Hoffman (1986) found that site differences accounted for an average of 19 percent of the variance in Army hands-on test scores. His analyses also indicated that statistical standardization procedures would control for these differences without unduly biasing the scores.

Scorers are human, and even the best trained will make mistakes in the administration and scoring of hands-on tests. The concern that active duty scorers will contribute more error to test scores than will retired military personnel is a hypothesis that has not been thoroughly tested. As described below, however, the Air Force reported high interscorer agreement for performance scores across all AFS, regardless of whether active duty or retired NCOs served as scorers (Hedge & Teachout, 1992), and the two AFS scored by retired NCOs did not exhibit higher ASVAB validities than the other AFS (Hedge, Teachout, & Laue, 1990).

A third potential contaminant to performance scores is one that is shared with all of the other criterion measurement methods. It is closely related to, but not exactly the same as the task experience issue described previously. In the discussion of relevance, we were concerned about situations in which an incumbent is tested on tasks that are not

part of his or her particular job. Here, we are concerned about performance differences that are due to differences in total job experience; that is, tenure. The Services tested personnel within a certain time-in-service limits, but this still allowed a range of experience levels. This range was large enough to permit significant covariation between performance scores and time-in-service to occur across all jobs included in the JPM project (OASD, 1989). Furthermore, Schmidt, Hunter, and Outerbridge (1986) have demonstrated that increased variance in job experience reduces the correlation between predictors and job performance criteria.

All of these factors probably contributed to score contamination in the JPM project, but it seems doubtful that they were critical factors in performance on most tasks tested. This assumption seems warranted given the validity estimates which indicate that ASVAB can be used to predict job performance criteria. It is also true, however, that the potential costs of these problems is significant and efforts to minimize their effects, both administratively and statistically, should continue.

Reliability. The Marine Corps was the only Service to have developed "parallel" forms of the performance tests. A sample of 86 infantrymen who took both sets of tests about one week apart yielded an alternate forms/test-retest reliability estimate of .83 (Mayberry, 1988). The Marine Corps also computed test retest reliability estimates, based on a 7-10 day interval, on a larger sample of 188 infantrymen (Carey, 1990). The overall estimate was .70, and scores went up an average of one full standard deviation. Given the assumption that examinees perform the tasks on which they are being tested, such a large practice effect is a bit disconcerting.

Median internal consistency estimates for each of the Services were as follows: Air Force .75; Army .85; Navy .81; and Marine Corps .87 (Wigdor & Green, 1991). The suitability of these estimates when applied to individual work sample tests is questionable because steps within tasks are generally not independent. Thus, internal consistency coefficients can be expected to overestimate the test's reliability.

The interrater agreement indices which have been reported by the Services have been quite high (Carey, 1990; Doyle & R.C. Campbell, 1990; Hedge & Teachout, 1992). Indeed, the estimates have been so high that Wigdor and Green (1991) report that there has been some speculation that scorers conspired to match their ratings. They conclude, however, that this probably does not account for the high estimates of interrater agreement. However, these estimates may also overestimate reliability because the go/no-go base rate for each task step is probably not 50-50. For example, if each rater has a go/no-go base rate of 80/20, the percent agreement expected by chance is 68%. If the base rate for each rater is 90/10, the percent agreement expected by chance is 82%. Most likely, the inflation is not quite so extreme but interrater agreement statistics must control for the base rate to be more clearly interpretable.

To the extent that there is a lack of agreement across scorers, it seems reasonable to expect that this might be the result of difficulties associated with observing performance of some tasks. For example, the Army tested the task "Put on a field pressure dressing." To score this task, the scorer must check the tightness of the bandage, and in the process of doing so, may loosen it. A shadow scorer who follows behind the original scorer, then, may find that the bandage is too loose. Thus one scorer

rates "go" and the other scorer rates "no go." Another example would be tasks that are performed in a confined space in which the presence of two scorers is problematic. Thus, even higher estimates of interscorer agreement might be obtained if tasks such as these are eliminated from the analyses.

Generalizability analyses have corroborated the stability of scorers in the performance testing process (Shavelson, Mayberry, Li, & Webb, 1990). These analyses have indicated, however, that there is a considerable amount of score variation attributable to tasks (Kraiger, 1990; Shavelson et al., 1990; Webb, Shavelson, Kim, & Chen, 1989). In other words, examinees are likely to be ranked differently depending upon the task being tested. This may be due to task experience differences described earlier as well as real differences in the nature of the tasks which contribute to this phenomenon. The end result is the same, however. Work sample tests should cover as many tasks as possible, and if it is infeasible to test a reasonable number of tasks, measurement reliability will be low.

Discriminability. Although the degree of discriminability in test scores varied across tasks, jobs, and Services, there was enough range in the scores overall to allow significant correlations with ASVAB to manifest themselves. So what if this had not been the case? Although it is possible to play with the difficulty level, and thereby the variability, of written tests, it is much harder to do this with performance tests. This is because it can be difficult, if not impossible to eliminate particularly easy or difficult performance steps from scoring. Thus, the choice may be between dropping a task from testing if it does not sufficiently discriminate or keeping it for the sake of measurement validity, but recognizing that it will reduce score variability overall.

Practicality. Work samples and simulations are perhaps the least practical of all the measurement methods. Although they are not particularly expensive to develop, they typically take inordinate amounts of time and resources to administer. Those who developed and administered the infantry tests for the Marine Corps estimated that hands-on tests cost two and one-half times written job knowledge tests (Felker et al., 1988). Clearly, this measurement method is a luxury in the sense that it is capable of providing highly reliable and valid performance data, but at a significant cost.

Simulations

Instead of using the equipment, materials, and information actually present in the job setting for assessment purposes, simulations make use of artificial props. Thus, all else being equal, simulations will exhibit somewhat less measurement fidelity than the work samples described previously.

There appear to be two major categories of simulations. The first makes use of equipment that may be some combination of computers and audio/visual aids. Examples include interactive video tests of mechanical troubleshooting tasks and computerized cockpit simulators which allow simulation of pilot tasks. The second major category comprises simulations based on people, paper, and ideas without the aid of equipment. These simulations are typified by standard assessment center exercises such as in-baskets, role-plays, and leaderless group discussions. Note that these two major categories of

simulations roughly correspond to the distinction made by Asher and Sciarrino (1974) between motor and verbal work sample tests.

Computer/Visual/Audio Aids

Simulators

The use of computerized simulators is common in occupational training that requires interaction with complex, expensive equipment. Civilian air traffic controllers receive radar training on simulated workstations which have the capability to replay live air traffic or create air traffic environments with prespecified levels of activity (e.g., number of aircraft, type of aircraft). As another example, the Army trains tank crewmen on gunnery tasks using the Unit Conduct of Fire Trainer (UCOFT) which is a high fidelity simulator of the M1 tank's optics system (Smith & Graham, 1987).

Although high fidelity simulators are generally quite expensive to design and produce, they offer a safe and effective avenue for technical training in both military and civilian occupations. Furthermore, there are a variety of commercially-developed training simulators on the market which allow even relatively small organizations to make use of this technology. Even if an organization has to develop its own simulator, however, the design and production costs may be significantly outweighed by the costs associated with training on operational equipment which is probably also expensive and in short supply as well.

Whereas the use of sophisticated simulators may be a reasonable option for training, using the same type of technology for selection and classification research will be a realistic option much less often. For one thing, the resources that an organization can devote to training can be expected to be considerably greater than those that the organization can devote to criterion measurement. Training devices can be used routinely for large numbers of people whereas criterion measures are likely to be perceived as having short-term, limited use. By the time the need for another validation study rolls around, the equipment that is being simulated is likely to have significantly changed anyway.

So what about using training simulators for criterion measurement? This would be possible under certain circumstances, but these circumstances are not often present. The first question is availability. Are there enough simulators which can be offlined from training needs to allow for their use in research? Are the simulators in the same location as the incumbents who need to be tested or can they be transported there? In the JPM project, testing was conducted all over the world, and most training simulators are only available at technical schools.

A second major question is problem difficulty. Because the simulators are used for training, the problems which can be reproduced may not reach a high enough level of difficulty to allow sufficient performance variation in an incumbent sample. In some cases, this problem could be addressed with software enhancements. Such enhancements, however, can be prohibitively expensive to make.

The way in which trainee performance is scored and recorded is a third major issue that is likely to be encountered, and the last discussed here. Theoretically, a computerized simulator could be designed to record all kinds of performance-related information, including reaction time, various choices made during the course of task performance, and the results of those choices. Often, however, training simulators do not make full use of potential recording capabilities. In fact, they may not permanently record performance information at all. The nature of the training setting, which usually includes one-on-one instruction, will often make it unnecessary to go to the additional expense of recording a lot of performance information. Furthermore, there may be no need to score performance for training purposes. Rather, performance is simply observed by a trainer, and used as a coaching tool. As with the problem difficulty issue, it may or may not be feasible to adapt an existing simulator to better accommodate the need to record and score performance.

Thus, training simulators offer the potential to serve as a source of criterion information, provided that they do not suffer from the common shortfalls outlined above. Of course, once a simulator is identified that is available for use, exhibits performance variability, and allows performance to be scored and reported, it must also satisfy the other standards required of a suitable criterion measure (e.g., reliability, validity). Given that computerized simulators vary widely in the extent to which they faithfully reproduce operational equipment and problem scenarios, validity of measurement is an important issue.

In the longitudinal portion of Project A, the Army used a training simulator to test a task that had previously been characterized as infeasible for hands-on testing. The task was "engage targets with an M16 rifle." Although the Marine Corps tested this task using live fire in its JPM effort, such an approach was not practical for the Army because the Army's project was larger in magnitude and logistically more complex. The simulator used was the Multipurpose Arcade Computer Simulator (MACS). The MACS equipment consists of a computer monitor, keyboard, and a demilitarized rifle with a light pen attached to its end. Targets (moving and stationary) are shown on the screen and the soldier shoots at them with the rifle/light pen. The portion of the program which appeared most suitable for experienced soldiers constituted the test. Although the MACS was more practical to use than a live fire exercise, it still posed significant challenges for the data collection effort. These included the expense and problems associated with shipping the equipment, problems with equipment reliability, and computer incompatibility issues.

A lower bound internal consistency reliability estimate for the MACS scores was estimated to be .74 for a sample of 985 soldiers (Jay Silva, personal communication, 13 October 1992). The uncorrected correlation between MACS scores and scores on the psychomotor composite predictor was .34. Although these results are promising, it would be helpful to know how performance on the MACS correlates with actual M16 shooting skill, especially given that the realism of the simulator (i.e., shooting at a computer monitor) can only be characterized as moderate at best. A more definitive understanding of the simulator's validity would help determine the usefulness of the data.

The Marine Corps used both a live fire exercise and two commercially-developed video firing tests to test marksmanship skills. [Note that the video tests have been

described as both potential predictor tests (Felker et al., 1988) and surrogate criterion measures (Carey, 1990).] The live fire exercise made use of pop-up targets on a firing range. Considerable difficulties were encountered during the first couple of weeks of testing due to reliability problems with the pop-up targets. Comparability of the firing exercise across the two testing locations was also an issue. With regard to the video tests, the simulations were not very realistic, but equipment reliability was not a significant problem. One test simulated a trap shoot and the other simulated a safari hunt. Together, they yielded a performance score that was internally consistent ($r=.82$), but which had relatively low test-retest reliability ($r=.63$ with 7-10 day interval). Correlations with the live fire test have not been published.

The foregoing discussion leads us to conclude training simulators will rarely be adequate to satisfy criterion measurement needs. The Navy's futile search for training simulators suitable for use in the JPM effort bears this out (Kidder, Nerison, & Laabs, 1987). In those cases in which suitable simulators are available, however, they are likely to add significantly to the quality of the validation research. One example provided above was the predictive validity of Army-developed psychomotor tests as demonstrated by their correlation with MACS scores. As another example, Smith and Graham (1987) validated both the experimental psychomotor and perceptual ability tests from the Army's Project A on a group of officer trainees. The officers were being trained on M1 tank gunnery skills using the UCOFT simulator. The researchers found a multiple correlation of .57 ($n=95$) between the predictors and criterion. Such a large estimate of predictive validity for this motor task would probably not have been found without a hands-on simulation as the criterion. Consider, as well, that simulators are generally only developed for training the most critical job tasks. Thus, a search for simulators should be an automatic step in the identification of criterion measures for any technical job.

Computer-Assisted and Stand-Alone Audio/Visual Aids

There is a wide range of sophistication in this category of assessment techniques. Computerized interactive videodisc technology permits relatively realistic modeling of task stimuli and responses to those stimuli (e.g., using touch screen capabilities). Videotape used without the interactive capability allows accurate portrayal of task stimuli, but the way in which the examinee indicates how he/she would handle the task must necessarily lack realism.

Blunt (1989) discussed the feasibility of using computer-assisted interactive testing as a surrogate for work samples in the DoD JPM effort, with particular emphasis on four of the Air Force JPM jobs. She provided a taxonomy of computer testing formats and described examples of each type of test. In addition, she described the stimulus and response characteristics that can be simulated for various types of tasks (e.g., administrative, psychomotor). She concluded that the utility of computer-assisted interactive testing should be evaluated on a case by case basis, and that it is most likely to be cost-effective if developed in conjunction with operational personnel programs (e.g., job training as discussed above).

The Navy developed interactive video tests to measure performance on Electronics technician and Electrician's mate tasks (G. Laabs, personal communication, 21 September 1992). Tasks tested were the same as those tested using work samples. In

this testing format, task stimuli were presented on videotape. Examinees indicated the actions that they would take by touching the videoseen. Their actions would then determine the way in which the task problem would unfold.

The Navy designed a Tailored Response Test (TRT) to collect performance data from Electronics technicians (G. Laabs, personal communication, 21 September 1992). The TRT presents task stimuli on videotape (Thornton, 1987). Following visual presentation of a problem, the examinee is given a written paragraph. The examinee indicates the correct response to the problem by crossing out words and phrases in the paragraph so that it describes the correct action. As with the interactive video tests, tasks selected for testing on the TRT were the same as those tested on the work samples.

Examples of less extravagant technology can be found in the civilian arena. Candidates for board certification in veterinary surgery are presented with videotape and slides which depict such stimuli as radiographs, lame animals, and gram stain tests (American College of Veterinary Surgeons, 1992). Examinees are given background information orally to supplement the visual aid, and are asked a series of questions regarding each case that is presented to them in this manner. In one part of the examination responses are made in writing in an open-ended format, and in another part of the examination responses are given orally.

As another example, there are several commercially-marketed video-based instruments for assessing supervisory skills. Typically, these instruments are developed using critical incident methodologies. SMEs generate examples of critical incidents which illustrate effective and ineffective job behaviors. The situations in which these behaviors occur form the basis for problem scenarios which are presented to examinees via videotaped scripted performances. To construct a multiple-choice response format, alternative behaviors that could be exhibited in response to the situation must be generated. This could be done by having individuals from the incumbent population indicate what they would do in response to each situation. Alternatively, SMEs could be asked to generate common responses to these situations. In either case, SMEs rate the effectiveness of all possible responses to form the basis for the scoring key. Responses which are not obviously correct or incorrect are selected to appear on the examination. Alternative responses may be acted out on video for the examinees or may be listed in writing.

It is indeed unfortunate that the Navy has not published more information about the innovative simulation techniques it tried out in the JPM research effort. It has been noted that correlations between the interactive video test scores and the hands-on test scores were relatively low (G. Laabs, personal communication, 21 September 1992). This may be due to practice effects, however, as experience on the hands-on tests appeared to increase performance on the video simulations. This phenomenon would serve to decrease the reliability of the scores. Interactive video testing has the potential for offering a more practical performance testing strategy compared to hands-on tests. Such a measurement system could exhibit relatively high measurement fidelity and positive psychometric characteristics in a self-contained computerized package transportable all over the world.

Assessment Center Exercises

Assessment center testing technology originated in the 1940s with World War II. It was first used widely in the civilian arena by AT&T starting in the 1950s. It is typically used for selection and promotion decisions. In most assessment centers, candidates rotate through a series of exercises which simulate job-related interpersonal and cognitive activities. Performance is scored by a cadre of trained assessors. Each exercise is evaluated by more than one assessor, and often each assessor observes the candidate perform more than one exercise.

Gaugler, Rosenthal, Thornton, and Bentson (1987) conducted a meta-analysis of 50 assessment center validation studies. They obtained a corrected mean of .37 and corrected variance of .017, suggesting that assessment centers tend to be valid predictors of job performance and job progress (i.e., promotability). A significant moderator of validity was whether managers or psychologists served as assessors, with psychologists providing more valid evaluations. Validity was also improved with the use of increasing numbers of assessment devices. Other variables, including length of assessor training, time spent integrating information, length of assessment center, and time between predictor and criterion measurement, did not appear to moderate validity estimates.

In the JPM project, assessment center exercises were not typically used because the tasks being tested were entry level technical tasks not requiring interactions with others. To the extent that role-players were used, they fulfilled simple functions (e.g., holding out an arm to get a field/pressure dressing). A major exception was the series of three supervisory role-plays developed by the Army for use with their second term longitudinal validation sample (J. P. Campbell & Zook, 1990b). Job analyses identified at least three important tasks that could be feasibly simulated using a single actor/scorer. The tasks were (1) conduct performance counseling, (2) conduct disciplinary counseling, and (3) conduct one-on-one training. Scenarios were devised for each of these tasks and scripts were written for the actors role-playing subordinates. The role-play scenarios are summarized in Figure 6. SMEs helped devise the role-plays and associated scoring system. Examinees were scored on a series of behaviors appropriate for each role-play (e.g., states purpose of the counseling session clearly and concisely) using explicitly-anchored, three-point scales. Civilian scorers were provided with extensive training to learn their roles and to become accustomed to the scoring system.

Analyses of the Army role-play data collected from 1009 second term soldiers in 1988-1989 have been conducted (C. H. Campbell & R. C. Campbell, 1990). The mean score across all three role-plays was 2.26 with a standard deviation of .42, indicating a reasonable degree of performance variability. The median one-rater reliability estimate across items in all three role-plays was .71. This estimate is not particularly high, but is not too bad for a single rater. Role-play scores did not correlate highly with two other measures of supervisory skills. The uncorrected correlation with peer and supervisor ratings was .17 and the correlation with a written test of supervisory skill was .12. The degree to which these different measures should be expected to correlate, however, is a matter for further study.

PERSONAL COUNSELING ROLE-PLAY SCENARIO

Supervisory Problem:

PFC Brown is exhibiting declining job performance and personal appearance. Recently, Brown wall locker was left unsecured. You have decided to counsel this soldier.

Subordinate Role:

- Solider is having difficulty adjusting to life in Korea and is experiencing financial problems.
- Reaction to counseling is initially defensive, but will calm down if not threatened. Will not discuss personal problems unless prodded.

DISCIPLINARY COUNSELING ROLE-PLAY SCENARIO

Supervisory Problem:

There is convincing evidence that PFC Smith lied to get out of coming to work today. This soldier has arrived late to work on several occasions and has been counseled for lying in the past. You have instructed Smith to come to your office immediately.

Subordinate Role:

- Soldier's work is generally up to standards, which seems to justify occasional "slacking off." Slept in to nurse a hangover and lied to cover up.
- Initial reaction to counseling in a very polite denial of lying.
- If supervisor insists, soldier admits guilt, then whines for leniency.

TRAINING ROLE-PLAY SCENARIO

Supervisory Problem:

The commander will be observing the unit practice formation in 30 minutes. PVT Martin, although highly motivated, is experiencing problems with the hand salute and about-face.

Subordinate Role:

- Feelings of embarrassment contribute to the soldier's clumsiness.
- Solider makes very specific mistakes.

Figure 6. Army Supervisory role-play scenarios.

Note. From Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A edited by J. P. Campbell and L. M. Zook, 1990, ARI-RR-1597, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences.

The Army administered its supervisory role-plays under testing conditions that were below the standards feasible for stationary assessment centers. Testing took place outside, inside, in the rain, in the heat, and so forth. There were many different role-players and they could not be monitored all of the time to ensure that they stayed true to their roles. Despite these difficulties, useful data were generated. Future criterion measurement efforts, especially if they are somewhat reduced in scope from the Army work, should consider the use of role-play exercises as a feasible approach to task measurement.

VI. VERBAL TESTS

An alternative to the high-fidelity simulation is some form of verbal test. A traditional written test consists of questions about task performance usually delivered in a multiple-choice response format. This type of test is relatively low in development, administration, and scoring costs. However, to the extent that written tests reflect the declarative knowledge determinant of performance, rather than performance itself, they are *not measures of performance*. We should also note that the military is something of a special case in this regard. That is, to perform one's job is to be "ready." If readiness is a goal, then demonstrating current knowledge may well be a performance behavior. More innovative types of written tests may also be developed which incorporate job-relevant problem scenarios and response modes that elicit higher-level problem-solving abilities. In this chapter we also discuss oral examinations as an alternative to the traditional paper-and-pencil multiple-choice test.

In general, a significant advantage of paper-and-pencil measures is that they can be designed to cover a wide range of job tasks. To the extent that written tests reflect the knowledge determinants rather than performance itself, they are not direct measures of performance. Another major advantage is that written instruments can generally be administered to large numbers of examinees at one time and scoring can usually be done mechanically. These advantages are not shared by oral examinations.

Types of Verbal Tests

Structured Response Format

Written tests, and occasionally oral tests, typically have a structured response format, usually in the form of multiple-choice response options. Scoring for these types of tests is item based. Scores may be reported at the total item level or for subsets of items categorized by content (e.g., task). The typical paper-and-pencil multiple-choice test has an answer key which can be easily and objectively verified by consulting a reference document (e.g., training manual, course textbook). These tests are fact-based and measure declarative knowledge (McCloy, 1990)

For test content which is based on tasks, it is possible to create written test items which model performance requirements rather than simply textbook-type knowledge requirements. Performance-based items create a brief scenario and ask the examinee to indicate what should be done to address the situation. More traditional types of items are more likely to focus on how or why things work the way they do. Further, performance-based items are designed to minimize verbal demands by making liberal use of pictures and drawings to illustrate the problem. Job knowledge tests developed for the JPM project generally fell into this category of instrument (Baker et al., 1988; Bentley, Ringenbach, & Augustin, 1989; C. H. Campbell, R. C. Campbell, Rumsey, & Edwards, 1986; Felker et al., 1988; Vineberg & Joyner, 1988). The Army also developed similar for-research-only school knowledge tests for 21 MOS involved in the JPM research. The school knowledge tests did not use performance-based items.

Performance-based items can also go one step further to reflect job requirements that require judgment. That is, these are problems in which the best answer is a matter of judgment rather than a clearcut matter of right or wrong. The video-based supervisory skill tests described in the last chapter illustrate this type of test. We include in this chapter, however, tests that have the same type of standardized response format, but which present the problem scenarios in brief written paragraphs rather than via videotaped acting. The Army developed such a test to assess supervisory skills in second term soldiers (J. P. Campbell & Zook, 1990b). A sample item from the Situational Judgment Test is provided in Figure 7. Note that examinees were asked to indicate which response option they believed to be most effective (M) and the option which they believed to be least effective (L).

You are a squad leader. Over the past several months you have noticed that one of the other squad leaders in your platoon hasn't been conducting his CTT training correctly. Although this hasn't seemed to affect the platoon yet, it looks like the platoon's marks for CTT will go down if he continues to conduct CTT training incorrectly. What should you do?	
<input type="text" value="L"/>	a. Do nothing since performance hasn't yet been affected.
<input type="text"/>	b. Have a squad leader meeting and tell the squad leader who has been conducting training improperly that you have noticed some problems with the way he is training his troops.
<input type="text"/>	c. Tell your platoon sergeant about the problem.
<input type="text" value="M"/>	d. Privately pull the squad leader aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

Figure 7. Army Situational Judgment Test Sample Item.

Note. From Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A edited by J. P. Campbell and L. M. Zook, 1990, ARI-RR-1597, Alexandria, VA: US Army Institute for the Behavioral and Social Sciences.

Structured response tests may also be adaptive in nature. Because job performance is multidimensional, however, adaptive testing technology based on item response theory is unlikely to be very useful. On the other hand, latent image testing has been used by a variety of licensure and certification testing programs and was experimented with by the Navy in its JPM effort (G. Laabs, personal communication, 21 September 1992). This type of testing makes use of a special inking process which allows the examinee to select an answer, then get additional information for the next question based on his or her answer to the preceding question. Perhaps the most difficult problem with latent image testing is the development of a workable and valid scoring system.

Unstructured Response Format

Verbal tests which do not have a highly structured response format may be scored as described above (i.e., on the basis of responses to items) or on the basis of how well responses to items demonstrate relevant knowledges, skills, and abilities. In either case, the scoring scheme is less objective than most standard multiple-choice tests.

Two types of written tests with unstructured response formats are essay tests and accomplishment records. Essay questions may either be knowledge-based (e.g., Compare and contrast...) or performance-based (Describe how you would...). Essay tests are desirable because they require examinees to generate answers rather than simply recognize the best answers from those provided. Disadvantages include difficulties associated with the development of scoring criteria and the training of scorers to reliably apply those criteria, and the fact that such tests rely heavily on writing skills that may not be job-relevant.

Like the performance ratings to be discussed in the next chapter, accomplishment records are not tests per se. Rather these instruments ask individuals to describe verifiable experiences they have had which demonstrate their job-related knowledges, skills, and abilities (Hough, 1984). Respondents are given detailed guidance on the types of experiences which are admissible and the level of detail that is required in their answers. To our knowledge, this assessment strategy has not been used for criterion measurement. This is probably because it requires a reasonably high degree of motivation on the part of the respondent. Further, this assessment strategy may not be very suitable for most technical tasks, but it could be a useful alternative measure of supervisory skills. Probably the most intractable problem with this measurement method, however, is opportunity bias. A supervisor may be very skilled, but may not have had the opportunity to demonstrate this skill under particularly difficult circumstances (e.g., open hostility among subordinates).

Oral interviews are commonly used as predictors, but have potential as criterion measures as well. Research has indicated that structured interviews containing questions based on high quality job analysis information (e.g., critical incidents) can form the basis for reliable, valid predictor measures (e.g., Campion, Pursell, & Brown, 1988; Janz, 1982). Common formats include the situational interview (Latham, Saari, Pursell, & Campion, 1980), the behavior description interview (Janz, 1982) and experiential interviews which amount to an oral accomplishment record.

As with the written judgment-based tests described previously, interviews do not normally have simple right or wrong answers. Responses, therefore, can be evaluated using many different strategies. One method is to construct benchmark (e.g., excellent, fair, and poor) answers for each question. Another method is to create behaviorally-based rating scales for each relevant KSA using behaviors that are apt to be elicited from the set of interview questions. In this method, the interviewee is rated on each KSA based on responses to all questions taken together. In the former method, the interviewee is rated on his/her response to each question. With regard to interviewers, it is reasonable to expect that the reliability and validity of the interview will increase with increased interviewer training and the use of multiple interviewers (i.e., use of a panel interview format).

The potential contribution of oral interviews for criterion measurement purposes, as with assessment center methodologies, has not been widely explored. The exception is the oral interview portion of the walk-through-performance-testing used by the Air Force, but the Air Force's use of equipment to simulate actual task performance in a show-and-tell fashion is conceptually distinct from the interview alternatives being described in this chapter. For some job requirements, it seems reasonable to expect that the interview measurement technologies described here might offer practical, reliable, and valid alternatives to other measurement strategies. For example, situational interviews could be used to assess responses to critical work conditions that cannot be adequately simulated through work sample tests and which require the use of an unstructured response format for valid measurement.

Evaluation of Verbal Tests

Relevance. Fact-based verbal tests are typically based on job (or training) analysis information which suggests that they cover relevant content. In the JPM project, tasks were selected for testing in written job knowledge tests in the same manner as they were selected for testing in the work sample mode. Thus, the tests were relevant to the extent that the test questions were faithful to task content requirements. Given that all of the Services wrote performance-based items, it is likely that the items were as realistic as possible given the constraints of a written, multiple-choice test format.

Comprehensiveness. Three of the Services restricted coverage of their written tests to tasks that had been tested in a hands-on mode. The Army, however, sampled approximately twice as many tasks in the written format as in the hands-on format. This allowed the testing of important tasks that were considered infeasible for hands-on testing, and more comprehensive coverage of the performance space because a larger number of tasks were sampled. The disadvantage of the Army approach was that fewer items could be included to cover each task. Reliability estimates, however, suggested that this strategy did not unduly limit measurement reliability (J. P. Campbell, 1988).

Susceptibility to contamination. Verbal tests are just that - "verbal." Many jobs, particularly those at lower skill levels, do not require incumbents to read, write, or talk about their jobs to any great degree. Tests which require significant amounts of reading, writing, or oral articulation, therefore, will put less verbal examinees at an inappropriate disadvantage. This contamination of the verbal factor can be minimized with the use of

short-phrased performance-based test items that use pictorial aids to reduce the number of words required to ask and answer the question.

Another possible source of contamination is that fact-based verbal questions may also require the examinee to display knowledge that is not required on the job. A medical specialist may never need to know why certain procedures alleviate shock, he/she just needs to know when and how to perform them. Asking incumbents to demonstrate an understanding of the "why," instead of the "how" in this case would introduce unwanted variance into criterion scores.

Reliability. Based on available estimates, most internal consistency reliability estimates for total job knowledge test scores were reasonably high. The Army reported split-half reliability estimates ranging from .82 to .89 and coefficient alphas ranging from .89 to .92 (J. P. Campbell, 1988). The Marine Corps reported coefficient alphas for written tests ranging from .87 to .90 for infantry MOS tests and a test-retest (7-10 days) estimate of .73 (Carey, 1990). The Navy reported low alpha estimates for Machinist Mate knowledge tests (Kroeker & Bearden, 1987). Estimates were computed separately for incumbents working in engine rooms and generator rooms and for two different types of items (equipment geography versus step recognition). They ranged from .24 to .52. Estimates reported by the Air Force are considerably lower because they are computed at the task level which means that they are based on relatively few items (in some cases, as few as two). Task-level coefficient alphas were computed using field test data (Bentley et al., 1989), and ranged from -.57 to .90. The field test estimates are based on small samples (n equals 25 to 43).

Although most JPM written knowledge tests appeared to exhibit acceptable internal consistency, not all of them did. This may indicate a problem with some individual tests; it is difficult to speculate without further information. As with work sample tests, differences in performance on different tasks, or steps within tasks, should not be entirely unexpected. Thus, internal consistency estimates should be interpreted with caution. Indeed, low point-biserial correlations which are used to eliminate items from single construct ability tests should not be used indiscriminately in the development of performance tests.

Discriminability. Three of the four written Air Force tests had a mean percent correct of approximately 60 (Hedge et al., 1990). The fourth test was substantially easier, with a mean percent correct of 74 percent. The nine Army tests ranged in difficulty between 56 and 70 percent correct (J. P. Campbell, 1988). The Marine Corps infantry tests were a bit harder, with mean percent correct ranging from 44 to 52 (Carey, 1990). Standard deviation estimates were uniformly about 10, plus or minus 2. Thus, the written tests appeared to provide reasonable discrimination across examinees.

Practicality. High quality, performance-based verbal tests can be challenging to develop because it is difficult to write good items and a relatively large number of pilot test examinees are usually needed to generate reliable item statistics which are needed in the development process. For medium to large scale uses, however, the effort is justified. Administration of written tests is exceedingly convenient and economical, as is scoring if the scoring key can be applied mechanically. Administration costs (in terms of resources

and convenience) increase dramatically for oral tests. Scoring costs also increase with any strategy that involves human judgment, rather than a purely mechanical approach.

Summary

Verbal tests have been praised for their economy and convenience and disparaged for their lack of realism because they can only assess declarative knowledge. The increasing sophistication of written and oral test strategies, however, is allowing some headway on the realism issue. The written knowledge tests developed in the JPM project were all performance-based. Performance-based items make liberal use of figures and pictures to depict task stimuli, and relate to how a task is performed more so than to why it is performed in a certain way. Items can also be written which pose complex technical or supervisory judgment problems (e.g., the Army's Situational Judgment Test). These tests are harder to develop because the development of an answer key requires expert judgment rather than reference to training documents or textbooks. Finally, we note that verbal tests can be used to cover a wider variety of tasks in less testing time and allow the depiction of many different task conditions. Unfortunately, the Services have not fully examined the utility of these tests given the JPM data, although some work has been reported (e.g., Carey, 1990). Given the general economy and feasibility of this measurement method, such examination should be conducted to the fullest extent that the data will allow.

VII. PERFORMANCE RATINGS

Theoretically, performance ratings have the potential for being the highest fidelity criterion measurement method available. A work sample, whether taken in the context of actual or simulated job performance, is still only a sample. It cannot capture an incumbent's ability to accommodate all job demands over a sustained period of time. Performance ratings, however, generally attempt to do just that. Ideally, ratings would be a reliable and valid assessment of actual performance on all significant job requirements.

Unfortunately, it appears that humans have a very difficult time serving as unbiased recorders and evaluators of behaviors. A rater's efforts may be hindered or helped depending upon such factors as the relationship between rater and ratee, similarities between rater and ratee, cognitive complexity of the rater, quality of the rating instrument, nature of the rating context (e.g., coaching versus for-research-only), and quality of rater training (Landy & Farr, 1980). Despite the many problems associated with ratings data, however, ratings have been the most commonly used criterion measure in selection and classification research (e.g., Pearlman, Schmidt, & Hunter, 1980; Nathan & Alexander, 1988), and exhibit psychometric properties that are at least no worse than alternative measures, in spite of their many sources of potential error.

Summary of Ratings Collected in JPM Project

Although each of the Services collected ratings in the JPM project, the extensiveness of their efforts varied widely. Figure 8 summarizes the performance rating instruments developed and administered by each of the Services. Both the Air Force and the Army had a variety of rating instruments, whereas the Navy and Marine Corps incorporated relatively few ratings measures into their research, and these were administered to only a subset of the jobs that were studied. The Army attempted to gather ratings from two supervisors and four peers for each soldier and the Air Force sought to gather ratings from one supervisor, up to three peers, and from incumbents themselves. In comparison, the Navy and Marine Corps collected ratings from a single supervisor only. With regard to rater training, all of the Services provided at least some instruction. Although details are spotty, it appears that the Air Force and Army raters were provided with the most extensive training. This face-to-face rater training focused on common rating errors, frame of reference issues, and privacy of the data, and importance of accurate data (Hedge et al., 1989; Pulakos & Borman, 1986). In other words, training was both instructional and motivational.

	Task-Based	Dimensional/Job-Specific	Global/Force-Wide
Supervisor	Army, Navy ^a Air Force	Army, Navy ^b Air Force, Marine Corps ^c	Army, Navy ^a Air Force, Marine Corps
Peer	Army Air Force	Army Air Force	Army Air Force
Self	Air Force	Air Force	Air Force

^aMachinist mate and Radioman ratings only

^bRadioman rating only

^cMechanical maintenance occupation only

Figure 8. Performance Ratings Collected in Joint-Service JPM Project

Rating Sources

Supervisors. Supervisors are the most common source of ratings data. They are probably viewed as the most obvious source of ratings data because they are generally expected to routinely evaluate subordinates' performance, and are experienced performing this function (Landy & Farr, 1980). Furthermore, supervisors (at least first level supervisors) usually have sufficient opportunity to observe performance, they have the requisite background to evaluate performance, and they are accustomed to serving as evaluators. Compared to ratings provided by peers and the incumbents themselves, supervisor ratings tend to be the most severe (Nathan & Alexander, 1988).

Peers. For many types of jobs, an individual's co-workers have the opportunity to observe his or her performance on a daily basis. Indeed in some jobs, co-workers see more day-to-day behavior than supervisors, and may be exposed to different behaviors as well. Because of this, several researchers have argued that peer raters provide a source of performance data that may overlap, but is not completely redundant with supervisors' (e.g., Borman, 1974; Landy & Farr, 1980). Indeed, it may be appropriate in some cases to develop different rating forms for peers and supervisors which reflect the different behaviors they have the opportunity to observe. Whereas there are numerous potential problems with using peer rating data for performance appraisal purposes (Latham & Wexley, 1981), these problems are minimized in the research setting. Furthermore, training can be used to help peers learn to evaluate performance even though this may be something that they do not have much experience doing.

M. M. Harris and Schaubroeck (1988) conducted a meta-analysis of the relationship between ratings generated by supervisors, peers, and incumbents. Analysis of 23 correlations between peers and supervisors yielded a mean correlation of .62 (standard deviation .24), corrected for measurement error. Thus, although the peer ratings tended to be more lenient than supervisor ratings, they were correlated more

highly than one would have expected based on prior nonquantitative reviews (e.g., Landy & Farr, 1980).

In the JPM project, correlations between peer and supervisor ratings were reported for the Air Force and Army data. Analyses reported by Toquam et al. (1988) indicate that Army job-specific dimension ratings made by peers and supervisors correlate at considerably different levels depending upon the dimension being rated. Similarly, correlations reported by Kraiger (1990) for Air Force data indicate that the correspondence between peer and supervisor ratings varied depending upon the type of rating instrument used and the job being examined. Correspondence was particularly low for task-level ratings.

Self. Although self ratings may be a useful component to a performance appraisal process, their utility as a source of criterion data is often problematic. The major problem appears to be extreme leniency in the ratings. For example, C. C. Hoffman, Nathan, & Holden (1991) compared the predictability of supervisor and self ratings of performance. Supervisor ratings were correlated .25 (uncorrected) with the cognitive ability predictor composite and self ratings were correlated .03. Comparison of the rating distributions indicated that self ratings were a full standard deviation higher than the supervisor ratings. Such extreme restriction in range could easily account for the negligible validity estimate. M. M. Harris and Schaubroeck (1988) found that mean self ratings were a full standard deviation higher than supervisor ratings and one-half standard deviation higher than peer ratings.

The pattern of exceptionally lenient self ratings was borne out in the Air Force JPM data base (Hedge, Ringenbach, Slattery, & Teachout, 1989). Further, the Air Force found evidence for Thorton's (1968) hypothesis that self ratings of poor performers are more inflated than self ratings of effective performers. None of the other Services collected self ratings for validation purposes. In its criterion measure field tests, the Army had peer raters rate themselves as a component of practice in the rater training program (Pulakos & Borman, 1986). This strategy did not appear to help rater accuracy or rating distributional characteristics so it was not used in subsequent data collections.

In addition to being more lenient than peer and supervisor ratings, self ratings also appear to be less highly correlated with the other two sources. M. M. Harris and Schaubroeck (1988) computed corrected mean correlations of .35 (standard deviation .11) with supervisor ratings and .36 (standard deviation .19) with peer ratings. Recall that the corresponding correlation between peer and supervisor ratings was .62. Again, the Air Force JPM results echoed these general findings (Kraiger, 1990).

Rating Scale Content

Task-Based. The Air Force, Army, and Navy developed task-based rating scales. The Army tasks were rated using a seven-point scale and the Air Force tasks were rated using a five-point adjectively-anchored scale (J. P. Campbell & Zook, 1990b; Hedge & Teachout, 1986). The Navy tasks were behaviorally anchored (Baker et al., 1988; Bearden, Wagner, & Simon, 1988). As far as we can tell, rated tasks were the same as those tasks that had been tested using work samples and/or verbal tests. An exception was a rating booklet developed by the Army for administration to soldiers in those

secondary (i.e., Batch Z) MOS for which no job-specific criterion measures were developed. It listed 11 common soldiering task areas (e.g., performing first aid on self and other casualties), and was anchored with adjectives (i.e., poor to excellent).

The Army chose not to include task ratings in the validation analyses because they exhibited poor distributional properties, relatively low interrater reliabilities, and an unclear factor structure (J. P. Campbell, 1988). Researchers concluded that raters simply did not have sufficient opportunity to observe performance on each specific task to provide reliable, valid ratings. High rates of not-observed responses corroborated this hypothesis, as did the fact that the common task area scales, which were written at a somewhat higher level of generality, were more reliably rated.

A novel task rating approach is being explored by the Air Force in a non-JPM research effort. It is rather unique because it calls for measuring job performance in terms of quantity rather than quality (or speed rather than accuracy). Specifically, an alternative performance measure referred to as productive capacity provides criterion scores on a time metric (Carpenter, Monaco, O'Mara, & Teachout, 1989; Faneuff, Valentine, Stone, Curry, & Hageman, 1990; Leighton et al., 1992). A productive capacity score for an individual on a given task is derived by dividing the fastest possible time in which the task can be performed by the time it takes the individual to perform the task (Leighton et al., 1992). The time it takes the individual to perform the task is estimated by his or her supervisor.

Research on the productive capacity measure is in the early stages. The first study on a single AFS showed only modest correlations between quality and quantity performance ratings, but did show a reasonable correlation between quantity estimates and ASVAB performance (Carpenter et al., 1989). Leighton et al. (1992) provided a preliminary report in a study designed to improve the validity and reliability of supervisor time estimates. This initial report described the data collection effort as being overly complex and resource-intensive. The data collected in that effort have not yet been analyzed. Once they are, we will learn more about the validity of the supervisors' judgments because airmen were actually timed performing the tasks on which they were rated. Although there is still considerably more research needed to determine the usefulness of this performance measurement strategy, it offers potential as an alternative to traditional quality-based measurement methods.

Dimensional/Job-Specific. Each of the Services constructed job-specific rating scales for some or all of the jobs that they studied. The Army developed Behaviorally Anchored Rating Scales (BARS) for the nine primary MOS in its research, and the Navy developed BARS for both jobs for which it developed performance ratings (Radioman and Machinist's Mate) (Baker et al., 1988; Bearden et al., 1988; J. P. Campbell & Zook, 1990b). The Air Force constructed AFS-specific dimension rating scales for each AFS included in its research. Dimensions were based on factor analysis of occupational survey data (Hedge & Teachout, 1986). Behavioral descriptions were used to anchor the rating scales for each dimension.

Global/Force-Wide. The Air Force constructed two global rating forms. One included eight dimensions of performance that are common across all AFS. The other was a two-item form tapping technical proficiency and interpersonal proficiency (Hedge

& Teachout, 1986). Behavioral descriptions were used to anchor each of these rating scales. The Navy used two different sets of Navy-wide rating dimensions (Baker et al., 1988; Bearden et al., 1988). Both sets of dimensions were defined and anchored behaviorally.

The Army developed an Army-Wide BARS which was administered to all MOS participating in its JPM research (J. P. Campbell & Zook, 1990b). The rating booklet included two global items which asked the rater to judge the soldier's potential as an NCO and his/her overall effectiveness. The Army also constructed a summated scale consisting of items describing behaviors that might be exhibited in a combat-related situation. This scale was developed in response to policy-makers' interest in getting some indication of how a soldier might be expected to perform under combat conditions. Raters completed the scale by indicating how likely or unlikely it would be that the soldier they were rating would behave as described in the item. For the administration of second term criterion measures in the longitudinal validation, this instrument was supplemented with ratings of actual combat performance in cases in which both rater and ratee had been deployed together for Operation Desert Shield/Storm.

The Marine Corps had a two item global rating form which it used for all MOS tested: (1) How much assistance does this Marine require to do his job, and (2) If your unit deployed tomorrow, would you want this Marine to deploy with you (Felker et al., 1988).

The force-wide dimensions used by each of the Services are summarized in Figure 9. Several dimensions (e.g., technical knowledge/skill, effort, and leadership) appear several times in one form or another. Differences appear to be less related to true differences across Services than to differences in the way in which the dimensions were generated by each developer.

Evaluation of Performance Ratings

Relevance. As with other criterion measurement methods, the relevance of a set of rating scales is dependent upon the quality of the job analysis information used to construct it. Because rating scales can easily cover a great deal of content, developers might be inclined to include more than they should. For example, one might include "ability to get along with co-workers" as an organization-wide rating dimension without considering that, for certain jobs, this dimension may not be relevant.

<p>Army-Wide Rating Dimensions (First term) (J. P. Campbell & Zook, 1990b)</p> <ul style="list-style-type: none"> • Technical knowledge/skill • Effort • Leadership • Following regulation & orders • Integrity • Military appearance • Self-development • Self-control • Physical fitness • Maintaining assigned equipment
<p>Air Force-Wide Rating Dimensions (Lance, Teachout, & Donnelly, 1992)</p> <ul style="list-style-type: none"> • Technical knowledge/skill • Initiative and effort • Leadership • Knowledge of and adherence to regulations • Integrity • Military appearance • Self-development • Self-control
<p>Navy Radioman (Baker et al., 1988) - general dimensions only</p> <ul style="list-style-type: none"> • Acquiring and using technical knowledge/keeping up-to-date • Conscientiousness, extra effort, and devotion to duty • Working with others • Maintaining living/work areas • Security mindedness • Safety mindedness
<p>Navy Machinist's Mates (Beardon et al., 1988) - general dimensions only</p> <ul style="list-style-type: none"> • Technical procedures • Adaptability/dedication • Safety • Mechanical aptitude/ability

Figure 9. Force-Wide Rating Dimensions Used in JPM Project.

Another threat to relevance is the content of rating scale anchors. Behavioral anchors are desirable, but their specificity can detract from the instrument's relevance if they include behaviors which are not appropriate for the job or dimension in question.

Comprehensiveness. Because rating scales can be constructed to capture performance at relatively global levels, this method provides the means to capture performance more comprehensively than any of the other measurement methods considered in this report. In addition to having the capability to cover performance content comprehensively, they also can be used to capture all three determinants of

performance postulated by the J. P. Campbell et al. (1992) performance model. Using Army Project A data, McCloy (1990) demonstrated that performance ratings were determined by declarative knowledge, procedural knowledge and skill, and motivation.

Susceptibility to contamination. Objective indices of on-the-job performance (e.g., production rate, sales volume) suffer from contamination resulting from faulty equipment, poor markets, lack of supervision, and so forth. Potentially, at least, a knowledgeable rater can take a variety of such extenuating circumstances into account when assessing the performance level of an individual. It is difficult to examine the extent to which raters are able to do this. On the other hand, it has been shown that raters tend to contribute to the contamination of ratings data in a number of ways. For example, there is evidence that raters' evaluations are affected by personal likeability, stereotype generalizations, carelessness, and so forth (e.g., Pulakos & Wexley, 1983; Terborg & Ilgen, 1975). Raters may also intentionally ignore standards depicted on the rating instrument in favor of their own idiosyncratic beliefs regarding effective performance (Pulakos, 1984).

Common rating errors such as halo, leniency, and central tendency also threaten validity to the extent that they mask true differences in job performance within and across ratees. These rating errors and the biases described in the previous paragraph are particularly problematic with operational performance ratings. Careful scale development and thorough rater training have been shown to alleviate these errors at least on a short term basis (Pulakos & Borman, 1986).

A rater who evaluates components of performance that he or she has had little or no opportunity to observe will provide ratings that are contaminated by halo, hearsay, or whatever other heuristic the rater uses as the basis for a rating. For this reason, rating scales should generally include a "cannot rate" option so that raters are not required to invent ratings.

Reliability. Kraiger's (1990) generalizability analyses of the Air Force JPM measures suggested that the ratings had acceptable levels of reliability when averaged across items/forms and rating sources. G coefficients were computed for each of the eight AFS tested, and most coefficients were greater than .70 when the ratings were averaged in this manner. The impact of averaging across rating sources was illustrated with validity analyses reported by Dickinson, Hedge, and Ballentine (1988). Peer ratings, which were averaged across multiple peer raters, correlated .32 with AFQT whereas supervisor ratings, which were based on a single rater, correlated only .17 with AFQT. Both correlations were corrected for range restriction.

Army analyses showed reasonable levels of interrater reliability, estimated with intraclass correlations, across all jobs for all except the task-level ratings (J. P. Campbell & Zook, 1990b). The Army-wide rating scale estimates were .65 for supervisor ratings and .58 for peer ratings. The MOS-specific rating scale estimates were .55 for supervisor ratings and .42 for peer ratings.

We were able to locate only internal consistency reliability estimates for Navy machinist mate task-level ratings (Kroeker & Bearden, 1987). Estimates ranged from

.90 to .97, across 14 tasks, depending upon rating source and job type (engine room or generator room).

Because ratings were collected from a single supervisor only, the only reliability estimate that could be computed by the Marine Corps was an internal consistency estimate based on the correlation between ratings on the two items included on the global scale. This correlation was .81 for the infantry jobs ($n=1,148$; Carey, 1990).

Discriminability. In addition to threatening validity, halo, leniency, and central tendency rating errors lead to rating distributions which lack discriminability. A major reason why for-research-only ratings are suitable for use as criterion measures whereas operational rating often are not is because rater training coupled with assurances of confidentiality appears to successfully counter these rating tendencies (Landy & Farr, 1980; Pulakos & Borman, 1986).

Practicality. Ratings are a very practical criterion measurement method, but short cuts that maximize economy and convenience may threaten the quality of the data. Specifically, careful scale development, provision of thorough face-to-face rater training, and collection of data from multiple raters all appear to influence data quality. In many cases, one or more of these requirements may be completely infeasible to meet (e.g., being able to obtain ratings from more than one rater). Thus, on a case-by-case basis, researchers must determine if the rating procedures they are able to establish will be effective enough to justify collecting ratings data. More specific guidance from the research literature would help make these decisions easier to make. For example, is face-to-face rater training always necessary for obtaining data of high enough quality for validation purposes?

Further Analysis of JPM Data

Significant contributions to the performance rating literature could be made with further examination of the JPM ratings data. For example, the Army collected two types of information related to each rater's ability to make ratings. On the first rating form, raters indicated the length of time they had worked with the ratee and at the end of each form, raters indicated the degree of confidence they had in the ratings that they were asked to make. The other Services may have collected similar data that would be useful for exploring issues related to rating accuracy within and across rater sources.

Some, if not all, of the Services have access to ASVAB scores and other computerized personnel file information on the individuals who provided performance ratings. This information could be used to explore rater characteristics that might influence rating accuracy and reliability, a line of research advocated by Landy and Farr (1980) and others.

The fact is that the JPM ratings data base, across the Services, is very rich in terms of the availability of large sample sizes (including women and minorities in many cases), multiple rating sources, a variety of rating formats, many types of rating dimensions and levels of specificity, and a host of related variables (e.g., cognitive ability, work sample, and job knowledge data). Several researchers have begun to mine these data in various ways (e.g., Borman, White, Pulakos, & Oppler, 1991; Vance, MacCallum,

Coover, & Hedge, 1988). Many large and small contributions to various lines of investigation in the literature on performance appraisal and to future efforts to use performance ratings for military research are still waiting to be made.

Summary

Conceptually, ratings would be the ideal measurement method because they are intended to capture typical on-the-job performance. Yet, it turns out that people are often not very good at making ratings, with the major problem being various types of criterion contamination. As with written tests, reported analyses of the JPM ratings have been very limited. The Army work suggests, however, that carefully collected for-research-only ratings from multiple raters and rater types yield reliable and valid performance information (Pulakos & Borman, 1986). The feasibility of this measurement method, along with encouraging research of this nature, argue for continued consideration of its utility especially in conjunction with other methods.

VIII. ARCHIVAL RECORDS

The last set of criterion measurement methods we will discuss in this report consists of archival indices of performance. These types of indices may be based on measurement methods described previously (e.g., supervisory ratings) or other more strictly objective indices of performance such as work volume or quality control measures. The advantages of archival indices of performance are that they involve minimal development cost and are usually inexpensive to collect.

The difficulty with the vast majority of archival performance measures is that they are almost invariably deficient and/or contaminated (Guion, 1965; Smith, 1976). Being deficient means that the measure provides only a partial picture of the worker's effectiveness on the job, leaving important aspects of the job unmeasured. Contamination of archival measures occurs when factors that affect how well a person does with respect to the measure are beyond the individual's control. The former problem is not insurmountable assuming that one is not relying on administrative measures as the sole source of criterion data. Criterion contamination, however, is more often an intractable and fatal flaw.

For the remainder of this chapter, we will review a number of administrative indices that have been evaluated for use as criterion measures in military settings. Comments regarding the reliability, validity, and other characteristics of these indices will be made within each section.

Supervisor ratings

As discussed in Chapter VII, ratings made for research purposes can provide reliable and valid criterion data. Ratings made for operational purposes, however, tend to be of limited utility for research purposes because of the severe contamination due to various types of rater errors (Pulakos & Borman, 1986).

Both the Army and the Marine Corps considered the use of operational supervisor ratings in their respective JPM projects. The Army did not use them because Enlisted Evaluation Reports (EERs) are not routinely collected for first tour soldiers and because EER ratings cluster at the maximum possible value (Riegelhaupt, Harris, & Sadacca, 1987).

The Marine Corps examined the utility of proficiency and conduct ratings which are routinely collected in training and approximately every six months in the field (Hiatt, 1986). The focus was primarily on the proficiency rating because of the Marine Corps' interest in can-do rather than will-do performance. Hiatt's analyses of operational ratings data indicated that ratings were heavily skewed, with the mean rating being approximately 4.5 on a scale that had a maximum rating of 5. Despite the heavy skewness of the ratings distribution, the field proficiency ratings exhibited meaningful correlations with related variables. Specifically, on a sample of 152 automotive mechanics, correlations (corrected for range restriction) were .27 with hands-on test performance, .33 with written job knowledge test performance, .52 with final training school grade, and .37 with the ASVAB mechanical maintenance composite. Hiatt

concluded that the proficiency ratings alone would not be suitable for assessing criterion performance, but that they could contribute to a composite measure of performance.

Average proficiency and conduct ratings were collected in the Marine Corps JPM effort. Analyses of the infantry data corroborated the findings of Hiatt (1986) (Carey, 1990). Corrected validity estimates using four ASVAB composites ranged from .30-.31 for proficiency ratings and from .20-.22 for conduct ratings.

Promotion rate

Although the Marine Corps collected promotion-related data on its mechanical maintenance MOS, the Army was the only Service to use promotion rate as a variable in its JPM validation analyses (Crafts et al., 1991; J. P. Campbell, 1987). The Army constructed a deviation-score promotion rate variable by comparing each soldier's pay grade with the average pay grade for soldiers in his/her MOS having the same time in service (Campbell, 1987). For second term soldiers, whether or not the soldier had ever been recommended for an accelerated promotion was also incorporated into the promotion rate score. The promotion rate variable comprised one element of the Personal Discipline performance factor in the Army validation analyses (J. P. Campbell & Zook, 1990b). As such, validity results with this variable used as the lone criterion have not been published. Its inclusion in the Personal Discipline composite score, however, indicates that it exhibited sufficient reliable variance to covary with other performance scores.

Training grades

The Air Force and Marine Corps obtained training grades for all enlistees included in the JPM research (Crafts et al., 1991; Felker et al., 1988). Citing problems with the quality and comparability of operational training grades across MOS, the Army developed for-research-only written multiple-choice tests of school knowledge for each MOS included in the JPM research (J. P. Campbell, 1987). These measures provided useful MOS-specific criterion information for over 20 jobs. The Navy did not collect training grade data for most ratings because of concerns similar to those articulated by the Army (G. Laabs, personal communication, 21 September 1992). Results of validation analyses using training criteria collected in the JPM project have not been published to date, although Army analyses will appear upcoming in an Career Force annual report.

Personnel File Records

The Army, Navy, and Marine Corps each evaluated the suitability of a wide array of variables available in enlistees' personnel records as criterion measures in the JPM effort (Crafts et al., 1991; Felker et al., 1988; Kidder et al., 1987; Riegelhaupt et al., 1987). The Navy's search was essentially limited to job-specific measures (e.g., Personnel and Training Education Program), and yielded sources that might be useful for helping to construct performance measures but no variables that would be directly suitable for inclusion in the JPM data base. Problems included lack of availability for many or most ratings, lack of currency of information, and limited accessibility.

The Marine Corps obtained physical fitness and rifle test scores from the personnel records of Marines in the infantry occupation. A larger array of variables were obtained for Marines in the mechanical maintenance occupation, including training courses, combat history, awards, number of positive and negative counseling sessions, and several indicators of disciplinary problems (e.g., punishments, courts martial). No analyses of these data, however, have been published.

The Army identified a number of variables that showed promise as performance indicators across all MOS. These variables included indicators of exceptional performance (i.e., awards, certificates of appreciation), indicators of disciplinary problems (i.e., Articles 15, flag actions), physical fitness test scores, marksmanship scores, and Skill Qualification Test (SQT) scores. Until recently, MOS-specific SQTs were routinely administered to monitor force readiness. The SQT scores used in the JPM project were based on written knowledge tests. In addition to these variables, the Army collected information on training courses and promotion board points for second term soldiers. The Army found that, by combining conceptually similar variables (e.g., all the indicators of disciplinary problems) to form composites, it could construct scores that exhibited acceptable base rates and distributions (J. P. Campbell & Zook, 1990b).

A significant problem for the Army was the accessibility of personnel record information. The information easily retrievable from central computerized records tended to be out-of-date. The up-to-date information available in each soldier's local paper personnel file was too resource intensive to gather on such a large-scale basis. In light of this problem, the Army examined the possibility of using a self-report form to collect this information directly from the soldiers in the JPM sample (J. P. Campbell & Zook, 1990b). Findings indicated that this strategy provided accurate, up-to-date information that was not distorted by the soldiers. The lack of distortion is likely due to the fact that the data were being collected for research purposes only.

Although the Air Force did not include personnel file types of performance indices in its JPM effort, this Service routinely administers Specialty Knowledge Tests (SKTs) which are similar to the Army's SQTs. SKTs are AFS-specific, 100-item, written multiple-choice tests (Weissmuller, Dittmar, & Phalen, 1989). New tests are written annually, and scores are a significant determinant of the Weighted Airman Promotion System. Although the continuous rewriting of these tests would make score standardization a significant problem, it is still conceivable that the SKTs could offer some amount of reliable and valid performance data for predictor validation purposes.

Production Indices

Production indices include objective measures of performance output, such as sales volume, number of products produced, and number of products rejected for poor quality. Such indices are notorious for the type of contamination described at the beginning of this chapter. For example, a military recruiter's success rate, in terms of number of applicants enlisted per month, is likely to be greatly influenced by characteristics of his or her market area (e.g., size and geographical dispersion of the youth population, community attitudes toward the military). To some extent, statistical adjustments can be applied to reduce the effects of such contaminants, but only if the contaminants can be appropriately quantified. The military also has to accommodate the

fact that most of its jobs are rather heterogeneous, if only because there are so many incumbents in so many locations. This situation is likely to make it even more difficult to identify production indices that would be suitable across incumbents.

None of the Services incorporated production indices into their JPM efforts. In addition to psychometric problems with these types of measures, they were probably also deterred by the fact that production measures will be available for only some types of jobs.

Attrition/turnover

In the military, turnover may be a result of attrition (i.e., failure to complete one's enlistment contract) or failure to reenlist after one's existing enlistment contract has expired. Because most first term enlistees are not permitted to reenlist, the Services are typically most interested in predicting attrition. The Army is the only Service to have examined turnover in its JPM research (White, Nord, & Mael, 1989). There are at least two reasons for this. First, an examination of turnover requires a longitudinal research design. Second, it seems unlikely that cognitive ability predictors would be strong predictors of attrition. In fact, the Army has focused its turnover research exclusively on its experimental temperament/biodata instrument.

When turnover is a criterion measure of interest, there are several problems commonly associated with its use that need to be considered. The first is a low base rate. Many enlisted personnel serve only one enlistment so turnover of this type has a high base rate. Yet, relatively few enlisted personnel fail to complete the terms of the first enlistment contract, making the base rate for this variable fairly low. In general, however, turnover base rates are typically higher in the military than in civilian organizations making this variable correspondingly more useful.

Researchers have distinguished between voluntary and involuntary turnover and between functional and dysfunctional turnover (Dalton, Krackhardt, & Porter, 1981). These categorizations attempt to distinguish turnover that is desirable from that which is not. Other researchers have distinguished unpredictable turnover from predictable turnover, where unpredictable turnover is that which is due to things out of the individual's control, such as medical or family problems (White et al., 1989). Turnover may be examined at finer levels, for example differentiating between turnover due to low satisfaction; medical, disciplinary, or family problems; and so forth. Of course one must be careful with this strategy because it could reduce base rates to unacceptable levels.

Most organizations maintain turnover records which include some indication of the reason for leaving. Indeed, the military maintains records which include dozens of reasons for leaving. The reported reason for leaving, however, may not be accurate due to a variety of factors (Campion, 1991). This presents another problem with the use of turnover as a criterion variable. For example, a manager may allow a poor performing employee the option of quitting so that his/her employment record is not marred with firing. Thus, care must be taken to evaluate the accuracy of turnover records. Identification of common "mistakes" could be useful in determining the best way to interpret the information for scoring purposes.

Summary

Indices such as turnover, disciplinary actions, awards, and promotion rate show considerable promise as supplemental criterion measures, but they do not provide job specific performance information. The major exception is training grades. Training grades are problematic, however, because the tests were developed and used for reasons other than predictor validation and the nature of the score distributions vary widely across jobs. Furthermore, most of the Services report having problems with the comparability and relative comprehensiveness of the training data bases across jobs. The Marine Corps no longer even maintains such a data base. The Army presented one potential solution to this problem by constructing for-research-only training tests. This solution, however, is an expensive one and it would probably be preferable to develop written job knowledge tests instead.

Another exception is the MOS-specific SQT scores used by the Army. These scores yielded very useful performance measurement information, but the SQT is no longer administered by the Army. The utility of the Air Force's SKTs has not been evaluated. Given that the validation of classification systems requires job specific performance information whereas the validation of selection systems does not (necessarily), it appears that administrative records are potentially quite useful for selection research, but much less so for classification.

IX. DISCUSSION AND CONCLUSIONS

This chapter considers the issues that govern the choices among alternative criterion measurement methods and casts them against the research objectives that seem most relevant for future classification research.

Choosing Among Alternative Criterion Measures: Conceptual Issues

It would be a mistake to assume that there is one best criterion measure, or conversely, that they are all equally useful. The major considerations that are relevant for making choices among criterion measures are outlined below. They progress from the general to the more specific; but even if the goals of measurement can be specified precisely, the conclusion must still be that multiple measures are appropriate.

Research vs. Appraisal

In general, collecting performance information for the purpose of operational performance appraisal is a very different context than collecting such data for research purposes, and the difference has very little to do with the specific theory of performance or measurement method being used (Ilgen & Feldman, 1983). It has everything to do with the reward-contingencies that operate on those who provide the data. For good and sufficient reasons it may not be a wise strategy, in the operational setting, to attempt to estimate an individual's true performance score as closely as possible. If you do that, people may never get promoted or it might play havoc with the salary structure. In the operational setting it also makes a difference if the appraisal is being done for evaluative (e.g., salary or promotion) or feedback (e.g., coaching) reasons. This fundamental difference between performance measurement for research and performance measurement for operational indicators as research criteria is a signal for caution when considering whether to use operational indicators as research criteria. They may produce very misleading estimates of predictive validity. In this regard, it is somewhat surprising that the administrative records used as criteria in the Army's Project A were as useful as they were.

The context of measuring performance for research purposes makes it more likely that the measurement system will try to estimate the individual's true performance score. There is everything to gain and little to lose by doing so. However, the specific sources of variation in the true score that should be allowed to operate or that should be controlled are a function of the nature of the research goal. For example, they would be different for selection/classification research, training evaluation, or the evaluation of alternative supervision/leadership strategies.

The General Research Objective

As noted previously, J. P. Campbell et al. (1992) have argued that for ongoing performance in a job setting there are three direct determinants of the variation in performance across people. Following the general literature in psychology these determinants are referred to as declarative knowledge, proceduralized knowledge and

skill, and motivation when "motivation" is defined as a combined effect from three choice behaviors: (1) choice to expend effort; (2) choice of level of effort to expend; and (3) choice to persist in the expenditure of that level of effort. These are the traditional representations for the direction, amplitude, and duration of volitional behavior. The important point is that the most meaningful way to talk about motivation as a direct determinant of behavior is as one or more of these three choices.

Accounting for individual differences in knowledge, skill, and choice behavior encompasses a very large number of research topics that will not be discussed here. From the trait perspective, almost a century of research has produced taxonomic models of abilities, personality, interests, and personal histories. Another major research tradition has focused on instructional treatment. At least three major types of such treatments are relevant in the job performance context - formal education, job relevant training (formal and informal), and previous experience. The possible antecedents of motivation, or choice behavior, are specified by the various theories of motivation. For example, an operant model stipulates that the reinforcement contingency is the most important determinant of the choices people make. Cognitive expectancy models say that certain specific thoughts (e.g., self-efficacy, instrumentality, valence) govern these three choices. Other models might see such choices as a function of certain stable predispositions such as the need for achievement. For example, perhaps certain kinds of people virtually always come to work on time and always work hard. A general schematic representing these points was shown previously as Figure 3.

A few general points should be noted. First, performance will not occur unless there is a choice to perform at some level of effort for some specified time. Consequently, motivation is always a determinant of performance and a relevant question for virtually any personnel selection problem is how much of the variance in choice behavior can be accounted for by stable predispositions measurable at the time of hire and how much is a function of the motivating properties of the situation, or the trait/situation interaction. Also, performance, that is not simply trial and error, cannot occur unless there is some threshold level of procedural skill. There may also be a very complex interaction between motivation and proceduralized knowledge and skill. For example, the higher the skill level, the greater the tendency to choose to perform, but skill level may have no relationship with the choice of effort level. That is, the three choices may be controlled by different antecedents.

Another reasonable assumption is that declarative knowledge is a prerequisite for procedural skill (Anderson, 1985). That is, before being able to use the procedural skills that are necessary for task performance one must know what should be done. This point is not without controversy (Nissen & Bullmer, 1987) and it may indeed be possible to master a skill without first acquiring the requisite declarative knowledge. Two examples that come to mind are modeling the social skills of your parents or modeling the "final form" of an expert skier without really "knowing" what you are trying to do. Nevertheless, given the current findings in cognitive research, the distinction between procedural skill and declarative knowledge is a meaningful one. Performance could suffer because procedural skill was never developed or because declarative knowledge was never acquired or because one or the other has decayed. Also, some data suggest that different abilities account for individual differences in declarative knowledge than account for individual differences in procedural skills (Ackerman, 1988). At this point,

the major implication is still that performance is directly determined only by some combination of these three elements.

Notice that there are now two levels of performance determinants, direct and indirect. Changes in selection/classification systems, changes in training and development programs, changes in management or leadership strategies, or other kinds of interventions are all indirect determinants that can only affect performance by changing one or more of the three direct determinants (i.e., knowledge, skill, the three choices). It is also true that in the real world there are a large number of physical or environmental constraints that could operate to restrict the range of performance differences among individuals. Conversely performance differences could be accentuated by applying the constraints differentially such as by giving some people better equipment than others. Such constraints are sources of contamination, not determinants of performance, and must be controlled for in some way.

If the research goal is to determine the validity of selection and classification procedures then the validation sample should be as homogeneous as possible in terms of the individuals' post selection experiences (e.g., training programs, quality of supervision, tenure) or in terms of the performance constraints under which each person must currently operate. Any systematic mean differences in true performance scores produced by differential experiences not under the individual's control constitute criterion contamination and for validation purposes downwardly bias the estimate of validity artifactually.

Also, the criterion measure should allow all three determinants of performance to operate. This should be reason enough to be wary of the standardized job sample as the sole source of criterion information. It purposely controls for much of the influence of the motivational determinant. However, the rank order of individuals may change a great deal when performance on the job sample is compared to performance on the very same tasks in the job setting. Perhaps the best available estimate of this latter correlation was obtained in the Sackett et al. (1988) study of supermarket checkout personnel. As luck would have it, the computerized recording system kept exactly the same scores on the same tasks over the course of a work day as were obtained from the standardized work sample. The reliabilities of both the job sample scores and real-time scores were high (.81-.85) but the correlation between them was much lower (.31). By comparison, for the Army Project A data, if the average correlation (across MOS) between the first tour hands-on job sample measures and the single rating of overall performance is corrected to the same level of reliability as the variables in the Sackett et al. study, the mean intercorrelation is about .30. That is, when corrected for differences in reliability, the correlation between standardized and naturally occurring job sample measures is about the same as the correlation between a standardized job sample and a single rating of overall performance.

If the research goal is not the validation of selection and classification procedures but the evaluation of the effectiveness of specific skills training programs, then the measurement goal is quite different. In such a context, the motivational determinants should be controlled. The research goal is to find out whether people have in fact mastered the skills. Choosing to actually use the skills on the job is a separate issue.

Lack of transfer should not be blamed on lack of mastery if the training program is in fact not guilty.

By similar reasoning, if the research goal is to evaluate the effects of new supervisory practices that are hypothesized to work because they influence subordinate motivation, then it is knowledge and skill that should be held constant by the measurement procedure. It would be undesirable to confound changes in subordinate motivation with changes in skill if the objective is to evaluate the supervisor as a "motivator."

The principal point here is that, so far as possible, the performance measures used as validation criteria should control for unwanted sources of variation and allow the relevant determinants to operate. As noted in the next section, even when the general goal is the validation of selection and classification procedures, the determinants of criterion variance that are designated as relevant can differ depending on the specific goals of the various parts of the selection and classification system.

The Specific Goals of Selection and Classification

In general, the choice of criterion measures for selection and classification research would depend on: (1) the goals to be maximized or minimized, (2) whether the decision process is single stage or multi-stage, and (3) the substantive nature of the differential requirements that exist across jobs, if any.

Certainly, attempting to maximize aggregate performance and attempting to minimize attrition are two different goals and they have different implications for criterion measurement. However, it is also possible to talk about maximizing performance on different components of performance or about minimizing different components of attrition. For some components the specific determinants may be different enough across jobs to make classification worthwhile (e.g., see Johnson & Zeidner, 1990), but for others it may not. This could be as true for attrition as it is for performance. Also, the nature of differential prediction across subgroups (e.g., race or gender) may be different for different performance components. None of these aspects of differential prediction can be detected unless the criterion measures used in research in fact allow them to operate. No matter what the classification potential of a new predictor battery, it most likely cannot be well demonstrated or conclusively documented with an overall global criterion measure.

If the decision process is a two stage system (i.e., an overall selection decision followed by classification), as it is in all the military services to some degree, then the relevant predictors and the appropriate criteria may be different for the two stages. For example, if a performance factor such as providing peer leadership or contributing to team performance was a function of essentially the same determinants across jobs, then predicted scores on such a criterion would not play a role in classification at stage II, but would be used for selection at stage I. If the between job differences are largely due to different knowledge and skill requirements, then a standardized job sample that attempts to hold motivation constant is a very appropriate measure. Further, if jobs are best distinguished by a few highly critical tasks then the criterion measures for evaluating

classification efficiency could be narrower in scope and could emphasize very reliable measurement of performance on the most critical tasks.

The questions of (a) how to make choices or tradeoffs among selection/classification goals (e.g., should the system emphasize increasing technical performance or reducing attrition, and to what degree?) or (b) how to combine information from multiple criterion components to make selection/classification decisions are really value judgments that will be made by default, if not explicitly. For example, if the prediction function for (a) demonstrating self control and avoiding discipline problems and (b) demonstrating high and consistent effort, do in fact differ, then the question of which outcome is to be emphasized to what degree when choosing criterion measures for selection/classification research is a management decision. Resolving the issue by asking for judgments of relative importance for each performance component and developing a weighted composite is one strategy that could be used (e.g., see Sadacca, J. P. Campbell, DiFazio, Schultz, & White, 1990). An alternative would be to cast the problem within the framework of conjoint measurement and to portray for the decision maker the actual tradeoffs (in terms of validity or prediction accuracy) that would accrue if one or the other component was emphasized. For example, if the prediction equation was chosen on the basis of predicting demonstrated effort, how well would that equation predict disciplinary problems? What would be the decrease in decision accuracy? What would be the gain in aggregate predicted scores on "demonstrations of effort" and what would be the loss in the predicted level of discipline problems. Again, to be able to assess the effects of either strategy there must be appropriate criterion measures of the relevant performance components.

Choosing Among Alternative Criterion Measures: Practical Issues

It is impossible to predict the particular constraints that will be imposed upon future enlisted personnel criterion measurement efforts. It is safe to say, however, that there will be limitations on both resources and time. Resource limitations are likely to include money, people, facilities, and/or equipment. Time limitations will have to accommodate the conceptualization, development, administration, and scoring of criterion measures. Aside from the ever-present Murphy's law, experience illustrates many problems that can be counted upon. For example, researchers planning to use criterion measures developed previously for a particular job may underestimate the amount of revision that is required to update the measures. This was the experience of Army researchers who updated hands-on and written job knowledge tests developed in 1984 for use in 1988. It can be very difficult to keep pace with rapidly changing technology, particularly when one's test is written and scored at a very specific level of detail.

There are many examples of how criterion measurement researchers have addressed the practical limitations imposed upon their efforts. Unfortunately, this is the type of information that rarely gets documented. An exception is the two reports documenting the development and administration of Marine Corps tests (Crafts et al., 1991; Felker et al., 1988). Each of these reports included a lessons learned chapter. More often, however, these lessons learned become part of the experiential knowledge possessed by those individuals who participated in the research, and are therefore easily lost. There is not very much that can be done to prevent this loss of experience, but

perhaps it would help to periodically remind ourselves as researchers that we need to seek out advice and insights from others who have done related work.

The Issue of Measurement Bias

Measurement bias is an issue that is often discussed in the context of predictor measurement, but rarely in the context of criterion measurement. The literature which is available is primarily related to performance ratings (e.g., Kraiger & Ford, 1985; Pulakos, White, Oppler, & Borman, 1989; Sackett & DuBois, 1991). Although criterion measures are not used to make selection decisions, criterion bias may lead to distortion of predictor validation findings. Furthermore, performance measures are often used as a basis for personnel decisions such as promotions and salary increases.

Part of the difficulty in studying criterion bias is determining an appropriate way to examine the data. As with predictor measures, a simple comparison of performance levels is not sufficient (although reporting of effect sizes would be informative). Oppler, J. P. Campbell, Pulakos, and Borman (1992) distinguished among three strategies for studying criterion bias: (1) total association approach, (2) direct effects approach, and (3) differential constructs approach. The total association approach determines the amount of variance in the criterion which is accounted for by subgroup membership. It says nothing, however, about the extent to which the variance accounted for by subgroup membership is valid or not or about the causal nature of subgroup score differences. In the direct effects approach, researchers attempt to hold true subgroup performance levels constant so that observed differences in performance scores can be attributed to bias. The differential constructs approach examines the relationship between scores on the target measure and other variables to identify subgroup differences in the apparent construct validity of the measure.

Oppler et al. (1992) examined the Army's Project A ratings data using all three of these analytic approaches. These researchers created a subset of the Project A data such that each incumbent had been rated by at least one black and one white supervisor, and by at least one black and one white peer. In addition, the data set included scores on hands-on and archival criterion measures. These data provided a unique analytic opportunity because of the array of variables included and the racial balance of the rater-ratee pairs.

- Point-biserial correlations between ratee race and five overall performance ratings were small, but the rater by ratee interaction was significant (Oppler et al., 1992). These "total association" findings are consistent with those of a large-scale civilian study by Sackett and DuBois (1991) but show much smaller effects than those estimated by Kraiger and Ford (1985), who conducted a meta-analysis of smaller-scale civilian studies. Thus, the common notion that whites rate whites higher than blacks and blacks rate blacks higher than whites may be an overgeneralization. To the extent that it does occur, at least in the Army data, peers are more inclined to follow this pattern than are supervisors.

Because of the need to hold subgroup performance levels constant, direct effects examinations of bias are often conducted in laboratory settings (Oppler, 1991). The real-world Oppler et al. (1992) direct effects analysis was made possible by controlling for

performance using scores from nonrating measures (training scores, hands-on scores, and archival measures). These analyses indicated that the total association estimates only slightly underestimated variance attributable to subgroup membership.

Examination of relationships among ratings and other Project A variables showed no appreciable evidence of differential construct validity for the ratings (Oppler et al., 1992). Taken in total, then, examination of the Army ratings data and other large scale studies suggest that ratings are not subject to appreciable levels of bias, particularly among supervisory raters. Examination of other criterion measures, however, has been quite limited. The three analytic alternatives distinguished by Oppler et al. and comprehensive literature reviews provided by Oppler (1992) and other authors (e.g., Kraiger & Ford, 1985) should help make the most of future research efforts.

Individual Contributions to Team Performance

The issue of team performance was raised in both the Task 1 and Task 3 reports of the Roadmap project (Knapp et al., 1992; Russell et al., 1992). In the Task 3 report, we discussed the difficulty of distinguishing between team performance and an individual's contribution to team performance, and the inappropriateness of using team performance indicators as criteria for selection and classification measures intended for use with individuals. With respect to the measurement of an individual's contribution to team performance, the difficulty of this distinction is no less apparent. Consider the use of a role-play exercise to measure contribution to team performance. The role-play would have to include several confederates and complex scripting to simulate teamwork in any meaningful way. The leaderless group discussion described in Chapter IV might serve as a reasonable alternative, but only if the exercise can be made to be reasonably consistent with actual job requirements. Although the verbal types of tests discussed in Chapter V do not offer much potential in this area, the Army and Air Force successfully used performance ratings to assess individual contributions to team performance in their JPM research.

Utility of JPM Instruments and Data

The Services took a large step forward in addressing the sparsity of research focusing on the "criterion problem" when they embarked upon the Joint-Service JPM project. The immediate goals of the project have been served with the submission of the final annual report to Congress (OASD, 1992) and the input of data into the DoD linkage/standard setting project (D. A. Harris et al., 1991). At this point, most of the Services have moved on to other projects and issues. That is understandable given the myriad of personnel research issues that have to be addressed with increasingly reduced staffing and resource levels. We also, believe, however, that a longer term view will illustrate the significant value of more fully utilizing the instruments and data generated through the JPM project to address the criterion problem. For example, what are the effects on estimates of classification efficiency of using total performance versus critical task performance as criteria?

To summarize the richness and scope of the project, consider that JPM criterion measures were developed for approximately 33 military jobs. Hands-on tests were developed for all of the jobs; written knowledge tests and rating scales were developed

for more than half of the jobs. In addition, simulations (e.g., interactive video tests) were developed for a subset of jobs, and administrative indices of performance (e.g., training grades) were identified for many of the jobs. Finally, two innovative types of relatively generic measures of first-level supervisory performance were developed by the Army. Using these instruments, criterion data were collected on over 15,400 enlisted personnel (26,400 if the Army's longitudinal validation sample is included).

At least in terms of published analyses, the Services have barely scratched the surface of these data. Furthermore, the development procedures and resulting criterion measures are bound to be useful for accelerating progress in future efforts, particularly if those future efforts are specifically designed to take advantage of this groundwork.

There are several steps that could be taken by the Services to ensure that the JPM work is used to maximum potential. One strategy is to document and disseminate the work that has already been completed. It was fairly difficult to gather many of the details that have been included in this report, and other details were excluded simply because we could not retrieve them. Furthermore, some of the information was documented, but only in the form of contractor reports that were not readily available (e.g., from the Defense Technical Information Center - DTIC). An ideal solution would be to create a central library of JPM and related research documents which would expand as more of the work gets documented. This is more than a convenience. It seems reasonable to expect that each of the Service's will be more likely to make documentation of previously-conducted work a priority if they believe that the information will be used to conserve future military fiscal resources. This will only be possible if the work is documented and readily available.

In addition to documenting work already completed, the Services need to ensure that the JPM data bases are available to researchers for additional examination. As we have mentioned several times already, we believe that there is a great deal more to be learned with these data. The data will not be used, however, unless they are accessible to those who have the time, resources, interest, and ideas needed to examine them. This is a sensitive issue for all concerned. Organizations that own the data are likely to feel somewhat proprietary about them. Individuals who were involved in the conceptualization, development, and administration of the measures are also understandably interested in analyzing the data. In the interest of progress, however, the Services need to allow those with "ownership" the time and resources to conduct these analyses and/or open the door further to others who can assist in the process of examining the JPM data more fully.

Finally, we would like to point out that the criterion measures themselves, as well as the training programs and documents associated with them are another valuable resource for future research activities. For example, some of the rating measures are Service-wide, making them suitable for jobs not included in the JPM research. Also, programs for training data collectors and performance raters should be useful for future data collection efforts.

Revisiting the Roadmap Objectives

As described in Chapter I, the Roadmap objectives identified in Task 1 which are most relevant to criterion measurement issues are as follows:

- Investigate criterion issues (e.g., How does the type of criterion used in validation affect estimates of classification efficiency and, ultimately, classification decisions? What is the appropriate criterion?).
- Investigate alternative selection and classification criterion measures in terms of their relative construct validity and susceptibility to subgroup bias.

The first objective is very broad as stated. We do not wish to reduce this breadth by trying to replace it with several individual objectives. The second statement was relatively focused to begin with and we have chosen not to suggest changes to it. Rather, it is our hope that the discussion provided in this last chapter has provided more specific direction to future efforts in these areas.

REFERENCES

- Ackerman, P.L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.
- Anderson, J.R. (1985). Cognitive psychology and its implications (2nd ed.). New York: W.H. Freeman.
- American College of Veterinary Surgeons (1992). The American College of Veterinary Surgeons Board Certification Examination Committee Handbook. Alexandria, VA: Human Resources Research Organization.
- Asher, J.J., & Sciarrino, J.A. (1974). Realistic work sample tests: A review. Personnel Psychology, 27, 519-533.
- Baker, H.G., & Blackhurst, J.L. (1986, November). Inter-technology transfer: Performance testing of jet engine mechanics. Paper presented at the 28th Annual Conference of the Military Testing Association, New London, CT.
- Baker, H.G., Ford, P., Doyle, J., Schultz, S., Hoffman, R.G., Lammlein, S., & Owens-Kurtz, C. (1988). Development of performance measures for the Navy Radioman (NPRDC TN 88-52). San Diego, CA: Navy Personnel Research and Development Center.
- Bearden, R., Wagner, M., & Simon, R. (1988). Developing behaviorally anchored rating scales for the machinist's mate rating (NPRDC TN 88-38). San Diego, CA: Navy Personnel Research and Development Center.
- Bentley, B.A., Ringenbach, K.L., & Augustin, J.W. (1989). Development of Army job knowledge tests for three Air Force specialties (AFHRL-TP-88-11). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Binning, J.F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74, 478-494.
- Blunt, J.H. (1989). Computerized performance testing as a surrogate job performance measure (AFHRL-TP-88-4). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Borman, W.C. (1974). The rating of individuals in organizations: An alternate approach. Organizational Behavior and Human Performance, 12, 105-124.

- Borman, W.C., White, L.A., Pulakos, E.D., & Oppler, S.H. (1991). Models of supervisory job performance ratings. Journal of Applied Psychology, 76, 863-872.
- Campbell, C.H., & Campbell, R.C. (1990, November). Army: Job performance measures for non-commissioned officers. Paper presented at the 32nd Annual Conference of the Military Testing Association, Orange Beach, AL.
- Campbell, C.H., Campbell, R.C., Rumsey, M.G., & Edwards, D.C. (1986). Development and field test of task-based MOS-specific criterion measures (ARI TR 717). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Campbell, C.H., Ford, P., Rumsey, M.G., Pulakos, E.D., Borman, W.C., Felker, D.B., de Vera, M.V., & Riegelhaupt, B.J. (1990). Development of multiple job performance measures in a representative sample of jobs. Personnel Psychology, 43, 277-300.
- Campbell, J.P. (Ed.) (1987). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1985 fiscal year (ARI TR 746). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Campbell, J.P. (Ed.) (1988). Improving the selection, classification, and utilization of Army enlisted personnel: Annual report, 1986 fiscal year (ARI TR 792). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Campbell, J.P., & Campbell, R.J. (1988). Productivity in organizations: New perspectives from industrial and organizational psychology. San Francisco: Jossey-Bass.
- Campbell, J.P., Dunnette, M.D., Lawler, E.E., & Weick, K.E. (1970). Managerial behavior, performance, and effectiveness. New York: McGraw-Hill.
- Campbell, J.P., McCloy, R.A., Oppler, S.H., & Sager, C.E. (1992). In N. Schmitt & W.C. Borman (Eds.) Frontiers in industrial/organizational psychology: Personnel selection. San Francisco: Jossey-Bass.
- Campbell, J.P., McHenry, J.J., & Wise, L.L. (1990). Modeling job performance in a population of jobs. Personnel Psychology, 43, 313-333.
- Campbell, J.P., & Zook, L.M. (Eds.) (1990a). Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel (ARI Technical Report 952). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.

- Campbell, J.P., & Zook, L.M. (Eds.) (1990b). Improving the selection, classification, and utilization of Army enlisted personnel: Final report on Project A (ARI RR 1597). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Campion, M.A. (1991). Meaning and measurement of turnover: Comparison of alternative measures and recommendations for research. Journal of Applied Psychology, 76, 199-212.
- Campion, M.A., Pursell, E.D., & Brown, B.K. (1988). Structured interviewing: Raising the psychometric properties of the employment interview. Personnel Psychology, 41, 25-42.
- Carey, N.B. (1990). An assessment of surrogates for hands-on tests: Selection standards and training needs (CRM 90-47). Alexandria, VA: Center for Naval Analyses.
- Carey, N.B. (1991). A comparison of hands-on and job-knowledge tests: Implications for better test development (CRM 90-201). Alexandria, VA: Center for Naval Analyses.
- Carpenter, M.A., Monaco, J.J., O'Mara, F.E., & Teachout, M.S. (1989). Time to job proficiency: A preliminary investigation of the effects of aptitude and experience on productive capacity (AFHRL-TP-88-17). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Childs, R.A., Oppler, S.H., & Peterson, N.G. (1992, August). Confirmatory analysis of two five-factor models of job performance. In M.G. Rumsey (Chair), Beyond generalizability of small r: Consistency of personnel research. Symposium conducted at the Annual Convention of the American Psychological Association, Washington, DC.
- Crafts, J.L., Bowler, E.C., Martin, M.F., Felker, D.B., Rose, A.M., Hilburn, B.G., & McGarvey, D.A. (1991). Develop and administer job performance measures for mechanical maintenance occupational area Volumes I and II: Test Development (AIR-70900-FR 02/91). Washington, DC: American Institutes for Research.
- Cronbach, L.J., Gleser, G.C., Rajaratnam, H. (1972). The dependability of behavioral measurements. New York: Wiley.
- Dalton, D.R., Krackhardt, D.M., & Porter, L.W. (1981). Functional turnover: An empirical assessment. Journal of Applied Psychology, 66, 716-721.
- Dickinson, T.L., Hedge, J.W., & Ballentine, R.D. (1988). Predictive efficiency of the ASVAB for the Air Force's job performance measurement system. In M.S. Lipscomb & J.W. Hedge (Eds.) Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-TP-87-58). Brooks AFB, TX: Air Force Human Resources Laboratory.

- Doyle, E.L., & Campbell, R.C. (1990, November). Navy: Hands-on and knowledge tests for the Navy radioman. Paper presented at the 32nd Annual Conference of the Military Testing Association, Orange Beach, AL.
- Dunnette, M.D. (1963). A note on the criterion. Journal of Applied Psychology, 47, 251-254.
- Felker, D.B., Crafts, J.L., Rose, A.M., Harnest, C.W., Edwards, D.S., Bowler, E.C., Rivkin, D.W., & McHenry, J.J. (1988). Developing job performance tests for the United States Marine Corps infantry occupational field (AIR-47500-FR 9/88). Washington, DC: American Institutes for Research.
- Faneuff, R.S., Valentine, L.D., Stone, B.M., Curry, G.L., & Hageman, D.C. (1990). Extending the time to proficiency model for simultaneous application to multiple jobs (AFHRL-TP-90-42). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Gaugler, B.B., Rosenthal, D.B., Thornton, G.C., & Bentson, C. (1987). Meta-analysis of assessment center validity. Journal of Applied Psychology, 72, 493-511.
- Guion, R.M. (1965). Personnel testing. New York: McGraw-Hill.
- Guion, R.M. (1979a). Principles of work sample testing: I. A non-empirical taxonomy of test uses (ARI TR-79-A8). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Guion, R.M. (1979b). Principles of work sample testing: I. Evaluation of personnel testing programs (ARI TR-79-A9). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Guion, R.M. (1979c). Principles of work sample testing: III. Construction and evaluation of work sample tests (ARI TR-79-A10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Guion, R.M. (1979d). Principles of work sample testing: IV. Generalizability (ARI TR-79-A11). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Harris, D.A. (1987, March). Job performance measurement and the Joint-Service project: An overview. In Proceedings of the Department of Defense/Educational Testing Conference on Job Performance Measurement Technologies. San Diego, CA.
- Harris, D.A., McCloy, R.A., Dempsey, J.R., Roth, C., Sackett, P.R., Hedges, L.V., Smith, D.A., & Hogan, P.F. (1991). Determining the relationship between recruit characteristics and job performance: A methodology and a model. (Final Report 90-17). Alexandria, VA: Human Resources Research Organization.

- Harris, M.M., & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. Personnel Psychology, 41, 43-62.
- Hedge, J.W. (1987). The methodology of walk-through performance testing. In J.W. Hedge & M.S. Lipscomb (Eds.) Walk-through performance testing: An innovative approach to work sample testing (AFHRL-TP-87-8). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Hedge, J.W., Dickinson, T.L., & Bierstedt, S.A. (1988). The use of videotape technology to train administrators of walk-through performance testing (AFHRL-TP-87-71). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Hedge, J.W., Ringenbach, K.L., Slattery, M., & Teachout, M.S. (1989, November). Leniency in performance ratings: Implications for the use of self-ratings. Paper presented at the 31st Annual Conference of the Military Testing Association, San Antonio, TX.
- Hedge, J.W., & Teachout, M.S. (1992). An interview approach to work sample criterion measurement. Journal of Applied Psychology, 77, 453-461.
- Hedge, J.W., & Teachout, M.S. (1986). Job performance measurement: A systematic program of research and development (AFHRL-TP-86-37). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Hedge, J.W., Teachout, M.S., & Laue, F.J. (1990). Interview testing as a work sample measure of job proficiency (AFHRL-TP-90-61). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Hiatt, C.M. (1986). Supervisor ratings analysis (CRC 537). Alexandria, VA: Center for Naval Analyses.
- Hoffman, R.G. (1986, November). Post differences in hands-on task tests. Paper presented at the 28th Annual Conference of the Military Testing Association, New London, CT.
- Hoffman, C.C., Nathan, B.R., & Holden, L.M. (1991). A comparison of validation criteria: Objective versus subjective performance measures and self- versus supervisor ratings. Personnel Psychology, 44, 601-618.
- Hough, L.M. (1984). Development and evaluation of the "accomplishment record" method of selecting and promoting professionals. Journal of Applied Psychology, 69, 135-146.
- Hunter, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. Journal of Vocational Behavior, 29, 340-362.

- Ilgen, D.R., & Feldman, J.M. (1983). Performance appraisal: A process focus. Research in Organizational Behavior, 5, 141-197.
- Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. Journal of Applied Psychology, 67, 577-580.
- Jenkins, J.G. (1946). Validity for what? Journal of Consulting Psychology, 10, 93-98.
- Johnson, C.D., & Zeidner, J. (1990). Classification utility: Measuring and improving benefits in matching personnel to jobs (IDA Paper P-2240). Alexandria, VA: Institute for Defense Analysis.
- Kanfer, R., & Ackerman, P.L. (1989). Motivation and cognitive abilities: An integrative-apptitude-treatment interaction approach to skill acquisition. Journal of Applied Psychology, 74, 657-690.
- Kavanaugh, M.J., Borman, W.C., Hedge, J.W., & Gould, R.B. (1986). Job performance measurement classification scheme for validation research in the military (AFHRL-TP-85-51). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Kidder, P.J., Nerison, R.M., & Laabs, G.J. (1987). Navy job performance measurement program: An examination of data bases, programs, and training simulators as sources of job performance information (NPRDC TN 87-28). San Diego, CA: Navy Personnel Research and Development Center.
- Knapp, D.J., Russell, T.L., & Campbell, J.P. (1993). Building a Joint-Service classification research roadmap: Job analysis methodologies (HumRRO IR-PRD-93-xx). Work performed under Contract No. F33615-91-C-0015 with the Air Force Armstrong Laboratory. Alexandria, VA: Human Resources Research Organization.
- Kraiger, K. (1989). Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project (AFHRL-TP-87-70). Brooks AFB, TX: Air Force Human Resources Laboratory.
- Kraiger, K. (1990). Generalizability of walk-through performance tests, job proficiency ratings, and job knowledge tests across eight Air Force specialties (AFHRL-TP-90-14). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
- Kraiger, K., & Ford, J.K. (1985). A meta-analysis of rater race effects in performance ratings. Journal of Applied Psychology, 70, 56-65.
- Kroeker, L., & Bearden, R. (1987). Predicting proficiency measures for machinist's mates. In Proceedings of the Department of Defense/Educational Testing Conference on Job Performance Measurement Technologies. San Diego, CA.

- Laabs, G.J., & Baker, H.G. (1989). Selection of critical tasks for Navy job performance measures. Military Psychology, 1, 3-16.
- Laabs, G.J., Berry, V.M., Vineberg, R., & Zimmerman, R. (1987, March). Comparing different procedures of task selection. In Proceedings of the Department of Defense/Educational Testing Conference on Job Performance Measurement Technologies. San Diego, CA.
- Lance, C.E., Teachout, M.S., & Donnelly, T.M. (1992). Specification of the criterion construct space: An application of hierarchical confirmatory factor analysis. Journal of Applied Psychology, 77, 437-452.
- Laue, F.J., Bentley, B.A., Bierstedt, S.A., & Molina, R. (1992). Data collection and administration procedures for the job performance measurement system (AL-R-1992-0118). Brooks AFB, TX: Armstrong Laboratory.
- Landy, F.J., & Farr, J.L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.
- Latham, G.P., Saari, L.M., Pursell, E.D., & Campion, M.A. (1980). The situational interview. Journal of Applied Psychology, 65, 422-427.
- Latham, G.P., & Wexley, K.N. (1981). Increasing productivity through performance appraisal. Reading, MA: Addison-Wesley.
- Leighton, D.L., Kageff, L.L., Mosher, G.P., Gribben, M.A., Faneuff, R.S., Demetriades, E.G., & Skinner, M.J. (1992). Measurement of productive capacity: A methodology for Air Force enlisted specialties (AL-TP-1992-0029). Brooks AFB, TX: Armstrong Laboratory.
- Lipscomb, M.S., & Dickinson, T.L., (1988). Test content selection. In M.S. Lipscomb & J.W. Hedge (Eds.) Job performance measurement: Topics in the performance measurement of Air Force enlisted personnel (AFHRL-RP-87-58). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
- Maier, M.H., & Hiatt, C.M. (1985). On the content and measurement validity of hands-on job performance tests (CRM 85-79). Alexandria, VA: Center for Naval Analyses.
- Mayberry, P.W. (1988). Interim results for the Marine Corps job performance measurement project (CRM 88-37). Alexandria, VA: Center for Naval Analyses.
- McCloy, R.A. (1990). A new model of job performance: An integration of measurement, prediction, and theory. Unpublished Ph.D. dissertation. University of Minnesota, Minneapolis, MN.

- McCloy, R.A. (1992). Methods for setting standards on predictor and criterion measures. In J.P. Campbell (Ed.) Building a Joint-Service classification research roadmap: Methodological issues (HumRRO IR-PRD-92-xx). Work performed under Contract No. F33615-91-C-0015 with the Air Force Armstrong Laboratory. Alexandria, VA: Human Resources Research Organization.
- McHenry, J.J., Hough, L.M., Toquam, J.L., Hanson, M.A., & Ashworth, S. (1990). Project A validity results: The relationship between predictor and criterion domains. Personnel Psychology, 43, 335-354.
- Murphy, K.R., (1987). Are we doing a good job measuring the wrong thing? In Proceedings of the Department of Defense/Educational Testing Conference on Job Performance Measurement Technologies. San Diego, CA.
- Nathan, B.R., & Alexander, R.A. (1988). A comparison of criteria for test validation: A meta-analytic investigation. Personnel Psychology, 41, 517-535.
- Nissen, M.J., & Bullmer, P. (1987). Attentional requirements of learning: Evidence from performance measures. Cognitive Psychology, 19, 1-32.
- Office of the Assistant Secretary of Defense (Force Management and Personnel) (1989, January). Joint-Service efforts to link enlistment standards to job performance: Recruit quality and military readiness. Report to the House Committee on Appropriations.
- Office of the Assistant Secretary of Defense (Force Management and Personnel) (1991, January). Joint-Service efforts to link military enlistment standards to job performance. Report to the House Committee on Appropriations.
- Office of the Assistant Secretary of Defense (Force Management and Personnel) (1992, April). Joint-Service efforts to link military enlistment standards to job performance. Report to the House Committee on Appropriations.
- Oppler, S.H. (1990). Three methodological approaches to the investigation of subgroup bias in performance measurement. Unpublished doctoral dissertation. University of Minnesota, Minneapolis, MN.
- Oppler, S.H., Campbell, J.P., Pulakos, E.D., & Borman, W.C. (1992). Three approaches to the investigation of subgroup bias in performance measurement: Review, results, and conclusions. Journal of Applied Psychology, 77, 201-217.
- Oppler, S.H., & Peterson, N.G. (1992, August). Comparing validity results using Project A concurrent and longitudinal samples. In M.G. Rumsey (Chair), Beyond generalizability of small r: Consistency of personnel research. Symposium conducted at the Annual Convention of the American Psychological Association, Washington, DC.

- Pearlman, K., Schmidt, F.L., & Hunter, J.E. (1980). Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. Journal of Applied Psychology, 68, 78-87.
- Pulakos, E.D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.
- Pulakos, E.D. (1986). The development of a training program to increase accuracy with different rating formats. Organizational Behavior and Human Decision Processes, 38, 76-91.
- Pulakos, E.D., & Borman, W.C. (Eds.) (1986). Development and field test of Army-wide rating scales and the rater orientation and training program (ARI Technical Report 716).
- Pulakos, E.D., & Wexley, K.N. (1983). The relationship among perceptual similarity, sex, and performance ratings in manager-subordinate dyads. Academy of Management Journal, 26, 330-342.
- Pulakos, E.D., White, L.A., Oppler, S.H., & Borman, W.C. (1989). Examination of race and sex effects on performance ratings. Journal of Applied Psychology, 74, 770-780.
- Riegelhaupt, B.J., Harris, C.D., & Sadacca, R. (1987). The development of administrative measures as indicator of soldier effectiveness (ARI Technical Report 754). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Russell, T.L., Knapp, D.J., & Campbell, J.P. (1992). Building a Joint-Service classification research roadmap: Defining research objectives (HumRRO IR-PRD-92-10). Work performed under Contract No. F33615-91-C-0015 with the Air Force Armstrong Laboratory. Alexandria, VA: Human Resources Research Organization.
- Sackett, P.R., & DuBois, C.L. (1991). Rater-ratee effects on performance evaluation: Challenging meta-analytic conclusions. Journal of Applied Psychology, 76, 873-877.
- Sackett, P.R., Zedeck, S., & Fogli, L. (1988). Relations between measures of typical and maximum job performance. Journal of Applied Psychology, 73, 482-486.
- Sadacca, R., Campbell, J.P., DeFazio, A.S., Schultz, S.R., & White, L.A. (1990). Scaling performance utility to enhance selection/classification decisions. Personnel Psychology, 43, 367-378.

- Schmidt, F.L., Hunter, J.E., & Outerbridge, A.N. (1986). The impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. Journal of Applied Psychology, 71, 432-439.
- Shavelson, R.J. (1991). Generalizability theory and military performance measurement: I. Individual performance. In A. Wigdor & B.F. Green (Eds.) Performance assessment in the workplace. Volume II technical issues. Washington, DC: National Academy Press.
- Shavelson, R.J., Mayberry, P.W., Li, W., & Webb, N.M. (1990). Generalizability of job performance measurements: Marine Corps rifleman. Military Psychology, 2, 129-144.
- Smith, P.C. (1976). The problem of criteria. In M.D. Dunnette (Ed.) Handbook of Industrial and Organizational Psychology (pp. 745-775). Chicago: Rand McNally College Publishing Company.
- Smith, E.P., & Graham, S.E. (1987). Validation of psychomotor and perceptual predictors of armor officer M-1 gunnery performance (ARI TR-766). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Steinberg, A.G., & Leaman, J.A. (1987, October). The Army leader requirements task analysis. Paper presented at the 29th Annual Conference of the Military Testing Association, Ottawa, Canada.
- Terborg, J.R., & Ilgen, D.R. (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. Organizational Behavior and Human Performance, 13, 352-376.
- Thornton, G.C. (1968). The relationship between supervisor and self appraisals of executive performance. Personnel Psychology, 21, 441-455.
- Thornton, R.F. (1987). Evaluation of a scannable method of scoring open-ended-response simulations. In Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies. San Diego, CA.
- Toquam, J.L., McHenry, J.J., Corpe, V.A., Rose, S.R., Lammlein, S.E., Kemery, E., Borman, W.C., Mendel, R., & Bosshardt, M.J. (1988). Development and field test of behaviorally anchored rating scales for nine MOS (ARI TR-776). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- Van Hemel, S., Alley, F., Baker, H.G., & Swirski, L.E. (1990, November). Job sample test for Navy fire controlman. Paper presented at the 32nd Annual Conference of the Military Testing Association, Orange Beach, AL.

- Vance, R.J., MacCallum, R.C., Coover, M.D., & Hedge, J.W. (1988). Construct validity of multiple job performance measures using confirmatory factor analysis. Journal of Applied Psychology, 73, 74-80.
- Vineberg, R., & Joyner, J.N. (1985). Development of an abstracted knowledge simulation of a hands-on test for machinist's mates (Final report for Battelle Contract DAAG29-18-D-0100). Alexandria, VA: Human Resources Research Organization.
- Wallace, S.R. (1965). Criteria for what? American Psychologist, 20, 411-417.
- Webb, R.J., Shavelson, R.J., Kim, K.S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurements: Navy machinist mates. Military Psychology, 1, 91-110.
- Weissmuller, J.J., Dittmar, M.J., & Phalen, W.J. (1989). Automated test outline development: Research findings (AFHRL TP-88-70). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.
- White, L.A., Gast, I.G., & Rumsey, M.G. (1986). Categories of leaders' behavior that influence the performance of enlisted soldiers (ARI RS-WP-86-1). Alexandria, VA: US Army Research Institute for the Behavioral and Social Sciences.
- White, L.A., Nord, R.D., & Mael, F.A. (1989, April). Setting enlistment standards on the ABLE to reduce attrition. Paper presented at the Army Science Conference, Durham, NC.
- Wigdor, A.K., & Green, B.F. (Eds.) (1986). Assessing the performance of enlisted personnel: Evaluation of a Joint-Service research project. Washington, DC: National Academy Press.
- Wigdor, A.K., & Green, B.F. (Eds.) (1991). Performance assessment for the workplace (Vol. I). Washington, DC: National Academy Press.
- Wolfe, J.H., Alderton, D.L., Cory, C.H., & Larson, G.E. (1987). Reliability and validity of new computerized ability tests. In Proceedings of the Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies. San Diego, CA.